Bargaining Over Taxes

Daniel Overbeck * Eliya Lungu†

Job Market Paper

January 9, 2025

[Click here for most recent version]

Abstract

This paper shows that bargaining over tax payments is an important feature of tax compliance and enforcement in lower income countries. Analyzing the universe of administrative tax filings from Zambia, we document sharp bunching in (i) dominated regions above tax schedule discontinuities, inconsistent with standard models of tax compliance and (ii) at round number tax payments, implying that certain payments are being targeted. Additional evidence from our own survey suggests that discussing tax payments with tax officials before filing taxes is widespread, in line with tax payments being the outcomes of bargaining. Such bargaining over taxes is consistent with fact (ii), as bargaining outcomes are often round and salient numbers, and with fact (i), because tax schedule discontinuities restrict the set of feasible bargaining outcomes. Finally, we generalize the conventional Allingham & Sandmo (1972) model to allow for bargaining as a mode of tax compliance. We show that bargaining leads to Pareto-improvements for both taxpayers and the state as long as state capacity is sufficiently low.

Keywords: Taxation, Bargaining, Development, Small-and Medium Enterprises, Tax Compliance, Zambia

*University of Mannheim, daniel.overbeck@uni-mannheim.de

†Zambia Revenue Authority

We thank Eckhard Janeba, Arthur Seibold, Nadine Riedel, Mazhar Waseem, Niels Johannesen, Doug Gollin, Justin Sydnor, Sebastian Seitz, Lukas Hack and seminar participants at the University of Mannheim, the 2024 NEUDC (Boston), the 2024 NTA (Detroit), the 2024 IIPF (Prague) and the 2024 German Development Economics Conference (Hannover) for useful comments and suggestions. Daniel Overbeck gratefully acknowledges financial support from the International Growth Centre for conducting the survey. The survey(-experiment) has been pre-registered with the AEA under the RCT ID AEARCTR-0012682 and received ethics approval from the Univeristy of Mannheim's Ethics Committee (EK 54/2023). The views expressed in this article are solely the responsibility of the authors, and should not be interpreted as reflecting the views of the Zambia Revenue Authority.

1 Introduction

In low- and middle income countries, state capacity is low (Besley and Persson, 2013). As a consequence, formal institutions often struggle to execute basic governmental functions effectively. Instead, informal institutions play an essential role. For example, credit markets, insurance systems as well as public good provision at the local level are partly organised through informal interactions between citizens and local elites or bureaucrats (e.g., Udry, 1994; Angelucci and De Giorgi, 2009; Olken and Singhal, 2011). Very little is known about the relevance of informal institutions on the tax side, however. As generating revenue through tax collection becomes increasingly important for developing countries (International Monetary Fund, 2024), it is crucial to understand informal institutions' role and potential in this process. In this paper, we study how informal institutions, and in particular, interactions between bureaucrats and citizens, shape tax compliance and enforcement *inside* the official tax system.

The setting we study is the taxation of small firms, which comprise the vast majority of taxpayers in most low-and middle income countries.[1] In particular, we study firms subject to turnover taxation in Zambia, a lower middle income and least developed country (United Nations, 2023). Analyzing administrative data on the universe of tax filings, we document two novel facts – both are inconsistent with standard models of tax compliance. First, firms bunch e.g. an excess number of firms locate in dominated regions above tax schedule discontinuities. Second, firms bunch at turnover amounts which imply round number tax payments (not necessarily round turnover). We further provide novel evidence from our own survey showing that discussing tax payments with tax officials before filing taxes is widespread, which is in line with tax payments being the outcomes of bargaining. Such bargaining over taxes rationalizes both empirical facts, as bargaining outcomes are often round and salient numbers, and because tax schedule discontinuities restrict the set of feasible bargaining outcomes. We provide several pieces of evidence, including a randomized survey experiment, to show that alternative explanations based on audit probabilities, optimization frictions, or mistakes cannot rationalize these bunching patterns. Finally, we generalize the conventional Allingham and Sandmo (1972) model of tax evasion to allow for bargaining as a mode of tax compliance. We show that, as long as state capacity is sufficiently low, bargaining over taxes leads to pareto-improvements for both taxpayers and the state.

[1] Small firms make up 80-90% of taxpayers and account for about 40% of GDP in the average developing country (World Bank, 2011, 2019).

Our empirical analysis builds on a novel and comprehensive dataset comprising the universe of more than 5.3 million turnover tax filings from 2015 until 2021 in Zambia. The data is at a monthly frequency and allows tracking firms over time and, importantly, across two major tax reforms in 2017 and 2019. Two key empirical patterns emerge. First, we measure bunching responses to tax schedule discontinuities. During the years 2017-2018, the Zambian turnover tax schedule features a linear tax rate and several tax brackets in which tax liability increases discretely by a fixed payment. Thus, net-of-tax turnover drops discretely at each bracket threshold. Such thresholds in the tax schedule, which induce drops in net-of-tax turnover imply that it is a strictly dominated choice to have (or report) turnover in a certain region above each threshold. Specifically, firms above tax bracket thresholds would be strictly better off when reducing turnover to below the threshold. Standard models of tax compliance would therefore predict bunching below the threshold but no mass above the threshold (Kleven and Waseem, 2013). In contrast, our data reveal strong bunching *above* each threshold. We find more than twice as many firms to report turnover just above the threshold compared to what our estimated counterfactual distribution predicts. The bunching is very sharp with most tax returns only exceeding the threshold by less than one Zambian currency unit, equivalent to around USD 0.05. Such bunching above thresholds is inconsistent with any standard preferences and thus a puzzling fact.[2]

Second, we find strong and sharp bunching at turnover amounts which imply round number tax liabilities, e.g., tax liabilities which are multiples of 10, 50, or 100. The Zambian setting is particularly appealing for this exercise as, before the tax brackets were introduced in 2017, the turnover tax schedule was flat at a rate of 3%. We can therefore clearly distinguish bunching at round number tax liabilities from the well-known phenomenon of bunching at round numbers of the taxable income itself (Kleven and Waseem, 2013; Carrillo et al., 2022). To see this, note that with a 3% tax rate, round liabilities often imply particularly odd amounts of turnover. Pooling all tax returns filed between 2015-2016, we document that more than 40% of the filed returns were such that the resulting liability was a multiple of 10. Importantly, these figures are not adjusted ex-post by the tax authority, but represent the raw numbers as appearing in the tax returns. This finding is inconsistent with standard models because there is neither an incentive to report nor any other reason for having such amounts of turnover. Instead, it suggests that the tax schedule gets *inverted* and certain payment amounts are targeted.

[2]Intuitively, standard preferences are such that utility is always increasing in consumption and that generating income is always costly (c.f. Saez, 2010; Kleven and Waseem, 2013; Kleven, 2016).

Exploiting the panel structure of the data, we show that the two empirical patterns are strongly connected. Firms that exhibit round payments prior to the reform have a much larger probability of bunching above thresholds than those that do not (or less so). This correlation is highly significant and robust to several alternative specifications. Leveraging administrative data on tax audits, we control for audit experience as well as various other characteristics and find that they neither change the significant effect of round payments nor have a significant effect themselves.

These empirical facts are puzzling as they are inconsistent with predictions from standard models of tax compliance. Thus, we present complementary results from a survey we conducted of 517 firms registered for turnover tax in Lusaka, the capital and largest city in Zambia. On average, the sample of surveyed firms matches the administrative data in terms of size, sector and gender. Several insights can be retrieved from the survey results.

The most striking result concerns the interactions between taxpayers and the Zambia Revenue Authority (ZRA). Nearly half of all respondents report to discuss the tax payments they are going to make with officials from the ZRA *before* filing their tax returns, or that such discussions are common. Of those, again half explicitly state that these discussions serve the purpose of finding agreements with officials on what should be paid. Put differently, firms bargain with officials over their tax payments. In line with the observations outlined above, the respondents state that bargaining evolves around the payment itself rather than the correct turnover. Furthermore, we estimate a significant and negative relationship with discussions over tax payments and the perceived probability of being formally audited for tax purposes. Bargaining over taxes could therefore serve as a preemptive measure to forego the formal procedure of filing and potentially being audited.

Motivated by these findings, we propose a theoretical framework of tax bargaining which can rationalize the empirical facts. We consider a risk averse firm choosing its tax payment conditional on its true tax liability (Allingham and Sandmo, 1972). The tax authority receives utility from tax revenues and penalty payments but incurs costs from auditing. We show that there exists a region of pareto-improving tax payments in this situation: relative to the standard non-cooperative case where audits induce risk for the taxpayer and costs for the tax authority, both parties are better off when agreeing on certain payments and not making any audits. This creates a potential surplus, which, in the model, is divided via Nash-bargaining.

The framework rationalizes both empirical facts. First, bargaining often leads to round-number outcomes, as round numbers can serve as focal points (Schelling, 1960; Janssen, 2006; Pope et al., 2015). In the model, we also show that the set

of payments which is bargained over is detached from a taxpayer's true liability, which is consistent with the notion that bargained payments often simply end up on round figures. Second, notches (i.e. discrete jumps in tax liability) effectively introduce regions of payments which cannot be reached anymore. This restricts the bargaining set and agreed upon payments accumulate just below but also just *above* the threshold. Importantly, our model also offers an explanation for why bargaining over taxes might be especially prevalent in less developed economies, with limited state capacity. A key feature of state capacity is the efficiency of tax audits. This efficiency crucially hinges on third-party reporting which is oftentimes non-existent for small businesses in less developed economies (Kleven et al., 2016). Our model demonstrates that as countries develop and build state capacity, the increased audit efficiency ultimately removes the scope for bargaining.

Our analyses include several additional pieces of evidence that rule out competing explanations for the observed bunching behavior. One prime concern might be that, in principle, the fear of being audited by the tax authority could incentivize firms to bunch above the threshold instead of below by itself, even in the absence of bargaining. Firms might simply trade a lower audit probability above the threshold against a larger tax payment. We address this concerns in four ways. First, we find no significant relationship between the event of being audited on whether a firm bunches above a threshold or not. Second, we find no substantial differences in empirical audit probabilities above versus below thresholds. Third, simulations from a standard model of tax evasion (Kleven et al., 2011; Allingham and Sandmo, 1972) show that even if perceived audit probabilities above- and below the threshold differ strongly, bunching above the threshold is highly unlikely to occur. We demonstrate that even under extreme assumptions where the probability of being audited jumps from 10% above to 30% below the threshold, only firms that evade at least 83% of their turnover would choose to stay above the threshold. Lastly, we run a randomized survey experiment to find that also shifting a firm's perceived audit probability upwards does not increase its stated propensity to bunch above a threshold. We provide further evidence to rule out optimization frictions or mistakes as an explanation for the observed bunching behaviour.

It is important to note, that the bargaining situations considered in this paper are distinct from mere corruption (Khan et al., 2016; Hindriks et al., 1999) where the tax collector receives bribes in exchange for lying. Bunching above thresholds– instead of below can hardly be rationalized with such collusion between taxpayer and tax collector. Both parties would always have a clear incentive to declare turnover below the threshold and share the difference in tax liability. Our survey also elicits that the

share of firms bribing tax collectors is much smaller than the share of firms engaging in bargaining. Finally, our model suggests that the government also benefits from bargaining through saving audit costs. Instead, we argue that incentive schemes for tax collectors may increase their effort in bargaining. We find that once a tax office hits its revenue target – and bonuses are being paid out to collectors – less firms tend to bunch above thresholds.

This paper contributes, first and foremost, to the literature on how tax administration is shaped by the institutional context of developing countries (e.g., Gordon and Li, 2009; Besley and Persson, 2014; Gadenne and Singhal, 2014; Okunogbe and Tourek, 2024). Among others, Olken and Singhal (2011) document that, for a large share of the population, tax collection and public service provision is organized entirely outside of formal institutions. Within formal institutions, recent studies have shown that engaging non-state actors such as local elites in tax collection (Balan et al., 2022) or subsidy targeting (Basurto et al., 2020) can overcome informational barriers and thus produce more efficient outcomes in low-income countries. Along these lines, Okunogbe and Pouliquen (2022) and Aman-Rana and Minaudier (2024) show that the digitization of tax collection and thus removal of personal interactions between taxpayers and tax collectors can partly lead to lower revenue collection. To the best of our knowledge, this paper is the first to show that such interactions can serve as a mechanism to determine tax payments through bargaining – an informal arrangement from which both parties, taxpayers as well as the tax authority, benefit. Aman-Rana et al. (2023) study another such informal arrangement, namely apparent corruption. In line with our interpretation, they argue that such arrangements constitute devices to overcome the low state capacity of governments in lower income countries. Regarding corrupt tax officials, Khan et al. (2016) and Hindriks et al. (1999) analyse situations where officials and taxpayers collude and lie about true tax liabilities. In this paper, we present empirical patterns that can hardly be rationalized by the presence of corruption. Instead, our finding of bunching above thresholds is consistent with bargaining even in the absence of corruption.

We further add to the broader literature on taxation and development from which, so far, particularly little is known about firm responses to taxation in low-income countries.[3] Most existing evidence is from Rwanda where it has been shown that a substantial share of registered firms is economically inactive (Mascagni et al., 2022) and many firms simply always file the same amount (Tourek, 2022). Relatedly, Almunia et al. (2023) document how firms in Uganda depart from alleged profit-

[3]Zambia switched frequently between low income and lower-middle income status according to the World Bank and only moved from low income to lower-middle income status in 2023.

maximizing behavior when filing taxes. We document similar albeit novel facts for the case of Zambia – another low income country – and offer bargaining over taxes as a yet underexplored explanation. Methodologically, our study contributes to and builds on work investigating how incentives are navigated by firms in low enforcement environments (Kleven and Waseem, 2013; Best et al., 2015; Anagol et al., 2022; Bachas and Soto, 2021). In particular, by contrasting the case of small firms in a least developed country to what is known about larger firms in countries at other stages of development, we highlight that the established predictions may not hold universally across the developing world.

The remainder of the paper is structured as follows. After section 2 details the institutional background, section 3 explains the data sources and empirical methodology. Section 4 shows the results from the administrative data and the survey. Section 6 rules out other explanations than bargaining. The theoretical framework is presented in Section 5. Section 7 concludes.

2 Background

Zambia is a lower middle-income country and classifies as a least developed country according to the United Nations (2023). In 2021, it had a population of 20 million people, a GDP-per-capita of USD 1137 PPP and a tax-to GDP ratio of about 16%. It thus closely resembles the average Sub Saharan African country along these dimensions (ATAF, 2018). Its tax-to-GDP ratio is low compared to the OECD country average of 34% and just above the ratio deemed necessary to meet basic needs of citizens and businesses (Gaspar et al., 2016).

Business taxation. In principle, every business in Zambia is required to be registered with the tax authority. However, as in all lower income countries, the majority of firms in Zambia is small and informal (not registered with the tax authority). Official statistics estimate that the informal sector accounted for nearly 90% of the country's employment in 2014 (Ministry of Labour and Social Security, 2018). Among the firms that are under the tax net, there is a crucial size distinction for how the tax base is determined. Businesses with annual turnover above ZMK 800,000 ($\approx$ USD $31,000$) are liable for the corporate income tax (CIT) where taxes apply to profits. Additionally, businesses are required to register for the Value-Added-Tax (VAT). If a business falls below this threshold it is liable for turnover tax in which, akin to a pure sales tax, taxes apply on turnover – at substantially lower rates.

Such systems are applied widely among lower income countries with the intention to simplify tax compliance by allowing for a simple measure of the tax base and thus tax liability.[4] More than 80% of Zambia's taxpayer population is registered only for turnover tax, highlighting that also the formal sector constitutes mainly of small firms. For firms under turnover tax, voluntary VAT registration is possible but very uncommon. Turnover taxes thus matter for a large share of the population and hence arguably for welfare. However, turnover taxes only account for less than 5% of total tax revenues.

It is common in Zambia to make tax payments in cash or per cheque. In 2016, more than 90% of tax payments were done in this way. Notably, this figure stood at about 20% in 2021, and thus, has decreased drastically throughout our study period. These numbers are aggregates for all tax types and likely to be larger for turnover tax payments (Zambia Revenue Authority, 2022).

Another crucial feature of tax compliance are personal interactions with tax officials. According to the World Bank Enterprise Surveys, more than 80% of small firms state to regularly visit tax officials in person. This ratio is slightly higher but comparable to other sub-saharan countries ($\sim 60\%$) but twice as large as the world average ($\sim 40\%$). On the other hand, corruption through tax collection is relatively low. Only 3.4% of small firms pay bribes to tax officials compared to the world average of 9.8%.[5]

Turnover tax reforms. The turnover tax in Zambia applies to monthly sales and therefore returns have to be filed and payments have to be made on a monthly basis. The system was introduced in 2009 and initially levied a tax rate of 3% on gross sales of firms with annual turnover below the CIT threshold. In 2017, the flat rate of 3% was replaced by a graduated bands schedule: turnover thresholds were introduced above which fixed payments had to be paid. Additionally, the turnover in excess of the threshold, was taxed at 3%. This created 7 tax brackets, described in Table 1.

Each threshold (except the first) implies a discrete jump in tax liability and thus in the average tax rate. For example, with a turnover of 4200, a firm had a tax liability of 36 ($= (0.03 \times (4200 - 3000))$. With 4200.01, the tax liability becomes 225. The average tax rate thus jumps from $\sim 1\%$ to $\sim 5\%$. Such discrete changes in the average tax rate are referred to as *notches* in the literature (Kleven and Waseem, 2013; Slemrod, 2013). Our empirical analysis exploits these notches to study firm

[4]Other examples of turnover tax systems in Africa include Nigeria, Kenya, Ghana, Uganda, Rwanda, Tanzania, Cameroon. Hoy et al. (2024) provide a recent overview of such systems in Africa.

[5]The data can be retrieved from `https://www.enterprisesurveys.org/en/data/exploreeconomies/2019/zambia`. Last accessed: June 12th, 2024.

Table 1: Turnover Tax Schedule 2017-2018

Turnover (in ZMK)	**Tax Liability (in ZMK)**
0 - 3000	0
3000 - 4,200	3% of monthly turnover above 3,000
4,200.01 - 8,300	225 per month + 3% of monthly turnover above 4,200
8,300.01 - 12,500	400 per month + 3% of monthly turnover above 8,300
12,500.01 - 16,500	575 per month + 3% of monthly turnover above 12,500
16,500.01 - 20,800	800 per month + 3% of monthly turnover above 16,500
Above 20,800	1,025 per month + 3% of monthly turnover above 20,800

Notes: This table depicts the turnover tax schedule which was in place in Zambia throughout 2017 and 2018. Tax liability increases discretely at 4200, 8300, 12500, 16500 and 20800. For example Liability is $36 = (4200 - 3000) \times 0.03$ at 4200 and 225 at 4200.01. This creates discontinuities in the budget set, referred to as notches.

responses to taxation. For this exercise, the Zambian turnover tax is especially appealing as neither the thresholds nor the payment amounts coincide with salient round numbers which could serve as psychologically appealing focal points. The reasoning behind the tax reform was that applying the same tax rate to firms of all sizes was perceived as regressive. In contrast, the tax schedule became partly regressive only after the tax reform. Fixed payments combined with a linear tax rate imply that the average tax rate is decreasing in turnover within each bracket. The tax reform therefore represented an ill-fated attempt to alleviate the concern of regressivity. In 2019, the schedule was again replaced by a flat schedule, this time with a rate of 4%.

3 Data & Methodology

In this section, we first explain the different data sources from both administrative records and the survey. Then, we describe the empirical methodology.

3.1 Data sources

Administrative data on tax returns. The administrative data used in this study has been provided by the ZRA and comprise the universe of turnover tax returns that have been filed in Zambia between 2015-2021. Each tax return contains information on the taxpayer identification number, total turnover amount, the tax liability, the tax office responsible for the taxpayer, the date a firm has registered for turnover tax and the period to which the tax return relates. The period is usually a calendar month. However in some cases, returns are covering multiple months. In total, the dataset includes more than 5.3 million firm-month observations. Additional sector

information is available for a smaller subset.

Administrative data on tax audits. We further observe audits under turnover tax which can be matched to taxpayers and the time periods to which they apply. While the reporting frequency for turnover tax is monthly, audits may apply to longer periods which means that multiple tax declarations are inspected. The audit data provides information on whether the audit resulted in payments of penalties or owed taxes.

Survey data. We complement the administrative data with a survey of Zambian firms. The survey data was collected throughout November and December of 2023.[6] A local team of six surveyors, hired through the survey provider Center for Evaluation and Development (C4ED), collected responses to 30-40 questions of 517 firms under turnover taxation in Lusaka, the capital city of Zambia. The interviews were based on a customized questionnaire and conducted in person. Access to information on taxpayer addresses is restricted by ZRA policy. Thus surveyors randomly approached small firms in Lusaka's main business areas to find survey participants. As questions on taxation can be sensitive, we believe an in-person approach to build a more trusting environment and thus to be better suited than phone interviews in this case (Blattman et al., 2016).

As survey respondents can not be linked to the administrative data, respondents were initially asked whether their business was registered for turnover tax or not and a number of questions served as checks as to whether their response was credible. Additionally, only if firms stated to have average monthly turnover of less than ZMK 70,000 ($\approx$ turnover tax threshold/12), the interview was continued. It was emphasized that the study was conducted independently of the ZRA and in fact, not even the surveyors were aware of any connection.

After the initial entry questions, firm characteristics such as sector, gender of the owner, number of employees and average turnover among others were collected. On average, survey sample firms are comparable to firms appearing in the administrative data. Further questions regarding accounting– and tax filing practices followed. The final part of the interview consisted of a randomized lab-in-the-field experiment, specifically designed to investigate potential channels driving the empirical facts established from the administrative data.

[6]The survey was designed and contracted by Daniel Overbeck alone, independently of Eliya Lungu or the ZRA.

3.2 Empirical approach

The empirical analyses of the administrative data largely rely on the estimation of *bunching* behavior (Kleven, 2016), a method to estimate local responses to certain thresholds. In particular, we compare the distribution of turnover tax returns observed in the data to a counterfactual distribution we estimate from a subsample of the data, omitting certain points of interest.

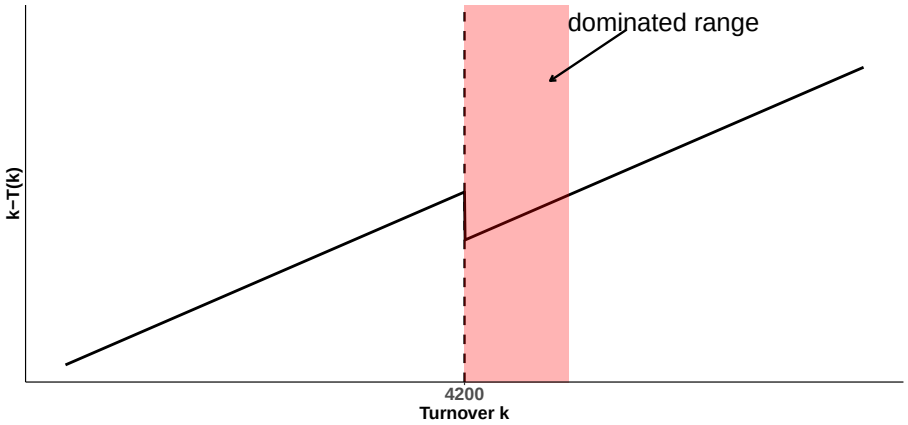
Bunching. We estimate two kinds of bunching behavior. The first kind we are interested in focuses on the tax bracket thresholds. More precisely, we investigate behaviour around the notches at which tax liability increases discretely. Through the lenses of conventional economic models, such discontinuities offer clear incentives for firms to behave in a certain way. Here, we focus on the simple intuition behind these incentives.[7] To understand the incentives, one should consider the budget set a firm is facing i.e. its net-of-tax turnover when the tax schedule features a notch. Panel (a) of Figure 1 sketches the budget set as a function of turnover around a notch akin to the ones in Table 1. Above the threshold, net of tax turnover drops and only exceeds net-of-tax turnover at the threshold when turnover is substantially larger. The red segment thus delineates a 'dominated range' in which a firm would be strictly better off with either lower or larger turnover. The incentives that are created by the notch are expected to result in a particular density of tax returns around the notch. Panel (b) of Figure 1 delineates this predicted density of tax returns (in the absence of optimization frictions). Firms are expected to be either reporting turnover just below the threshold or substantially above the threshold. As no firm would have turnover in the dominated region, we would expect a hole in the distribution just above the threshold.

To now estimate the extent of bunching below the threshold we proceed as follows. We coarsen the turnover data into bins of width 100 ZMK and draw on the approach developed by Kleven and Waseem (2013), in which we control for the *affected* turnover range, in which we would expect either excess mass (bunching) or missing mass (too little tax returns). To also account for potential bunching at round turnover as well as round liability amounts, we include dummies for both

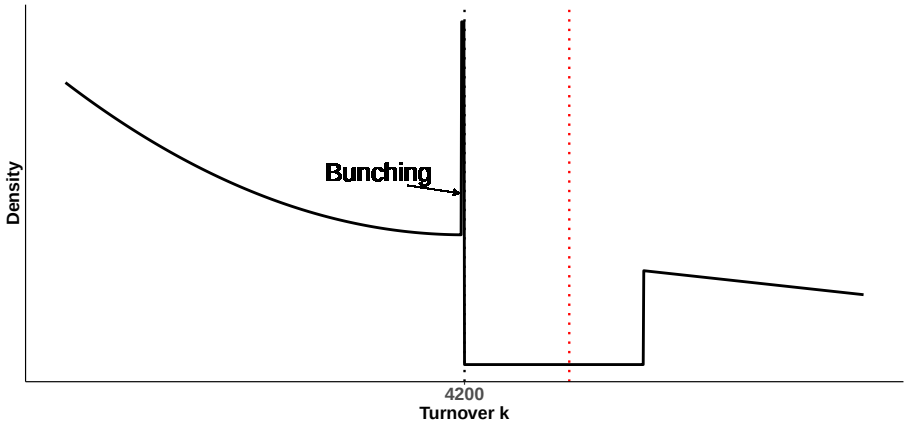
[7] Thorough theoretical treatments of bunching at notches are available in Kleven and Waseem (2013), Bachas and Soto (2021) or Slemrod (2013)

Figure 1: Bunching concept

(a) Notch in the budget set



(b) Expected excess- and missing mass



Notes: This figure illustrates the concept of bunching at notches. Panel (a) plots the budget set (i.e. net of tax turnover) on the vertical axis and turnover on the horizontal axis. The budget set is linear before the threshold, then drops discretely at the threshold and continues linearly from there on constituting. This *notch* establishes a dominated range (colored in red) above the threshold where net-of-tax turnover can be increased by either lowering or increasing turnover. Based on this tax schedule, panel (b) plots the expected distribution of firms around the threshold. One expects an accumulation i.e. *bunching* of firms just below the threshold and zero mass above the threshold within the dominated range.

cases. The estimating equation reads

$$\begin{aligned} c_j = &\sum_{i=0}^{p} \beta_j (z_j)^i + \sum_{i=z_L}^{z_U} \gamma_i \mathbf{1}[z_j = i] + \sum_{n \in \mathbf{N}} \xi_n \mathbf{1}\left[\exists k \in j | \frac{T(k)}{50} = n\right] \\ &+ \sum_{r \in \{100,500,1000\}} \rho_r \mathbf{1}\left[\frac{z_j}{r} \in \mathbf{N}\right] + \eta_j \end{aligned} \tag{1}$$

where j refers to a certain bin (i.e. $(4100, 4200]$). c_j and z_j capture the density (i.e. number of tax returns) and the upper bound of bin j (i.e. $z_j = 320$). The equation estimates a polynomial approximation of the distribution of tax returns

across turnover bins. β_j denote the coefficients of the polynomial vector. The coefficients ξ_n and ρ_r account for the influence of round numbers on both the liability as well as the turnover level. Turnover (i.e. the tax base) is denoted by k and the tax schedule is denoted by $T()$. The two indicator functions represent dummies for whether a turnover amount within a given bin coincides with a tax liability which is divisible by 10 and whether an upper bound of the bin itself coincides with a round number, respectively. z_L and z_U denote the lower and upper bound of the affected region, respectively. These parameters are chosen in an iterative procedure. We start with $z_L = z_U$ and increase z_U until the area above the counterfactual (excess mass in the data) equals the are between data and counterfactual (missing mass). The counterfactual distribution is retrieved from estimating Eq. (1) and then fitting $\hat{c}_j$ while excluding $\hat{\gamma}_i$ (i.e. the effect of being in the affected region). The counterfactual in this case therefore estimates the density of tax returns in the hypothetical *absence of a notch*. The extent of bunching below the threshold is finally estimated by

$$B_{below} = \frac{\sum_{j=z_L}^{\bar{z}} c_j}{\frac{1}{|j \in [z_L, \bar{z}]|} \sum_{j=z_L}^{\bar{z}} \hat{c}_j} \tag{2}$$

To illustrate that the canonical model (and with it its predictions) might not be applicable, we estimate bunching responses also *above* the threshold. This is done by comparing the counterfactual distribution as estimated by Eq. (1) in the first bin above the threshold with the observed data mass in the same bin.

$$B_{above} = \frac{c_{\bar{z}+100}}{\hat{c}_{\bar{z}+100}} \tag{3}$$

If the canonical frictionless model is applicable, then $B_{above} = 0$. If there are optimization frictions B_{above} may be larger than 0 but less than 1 (Kleven and Waseem, 2013). We calculate standard errors on both bunching numbers as well as z_U can by bootstrapping the residuals in Eq. (1).

The second kind of bunching revolves around *round number tax liabilities*. As is well documented in the literature, tax data, especially from low-and middle income countries often features bunching at round numbers. The round number focus has thus far been on the tax base level and is usually attributed to poor record-keeping by taxpayers (Kleven and Waseem, 2013; Carrillo et al., 2022). In this study, we estimate bunching at round number tax liabilities instead. With a tax rate of 3%, the Zambian setting allows us to distinguish between bunching at round liabilities and round turnover amounts. To see this, one can consider a firm desiring to pay

(for some reason) 100. The amount of turnover it needs to report is $\frac{100}{0.03} = 3333.33$. Clearly, neither lazy reporting nor firms actually having such a turnover and declaring it truthfully could explain an accumulation of observations at such odd amounts.[8] To get a sense of whether firms are in fact focusing on round liability points we estimate bunching at such odd turnover amounts. We do so by binning the distribution of tax returns into bins of width 10 ZMK and estimating the following equation:

$$c_j = \sum_{i=0}^{p} \beta_j (z_j)^i + \sum_{n \in \mathbf{N}} \xi_n \mathbf{1}\left[\exists k \in j | \frac{T(k)}{10} = n\right] + \sum_{r \in \{100,500,1000\}} \rho_r \mathbf{1}\left[\frac{z_j}{r} \in \mathbf{N}\right] + \eta_j \quad (4)$$

where the variables are defined as in Eq. (1).

We estimate the extent of bunching at round number tax liabilities by first fitting $\hat{c}_j$ as estimated by Eq. (4) while excluding the $\hat{\xi}_n$, but including $\hat{\rho}_r$. We then compare this fitted density $\hat{c}_j$ to the observed density c_j to calculate a bunching coefficient for a round number liability r in bin j:

$$B_r = \frac{c_j}{\hat{c}_j}. \quad (5)$$

As above one can calculate standard errors to these estimates by bootstrapping the residuals from Eq. (4).

4 Empirical facts

In this section, we apply our empirical methodology to the administrative tax data and establish two stylized facts, which are at odds with predictions from standard models of tax compliance. Furthermore, we present evidence from our own survey, describing how taxation works for small firms in Zambia.

4.1 Facts from the administrative data

Bunching below and *above* notches. We begin by estimating bunching responses to the notches in the tax schedule, which were introduced in 2017 (cf. Table 1). For our estimation we pool all tax returns that have been filed between January 2017 and December 2018 and thereby cover the whole time period the schedule was

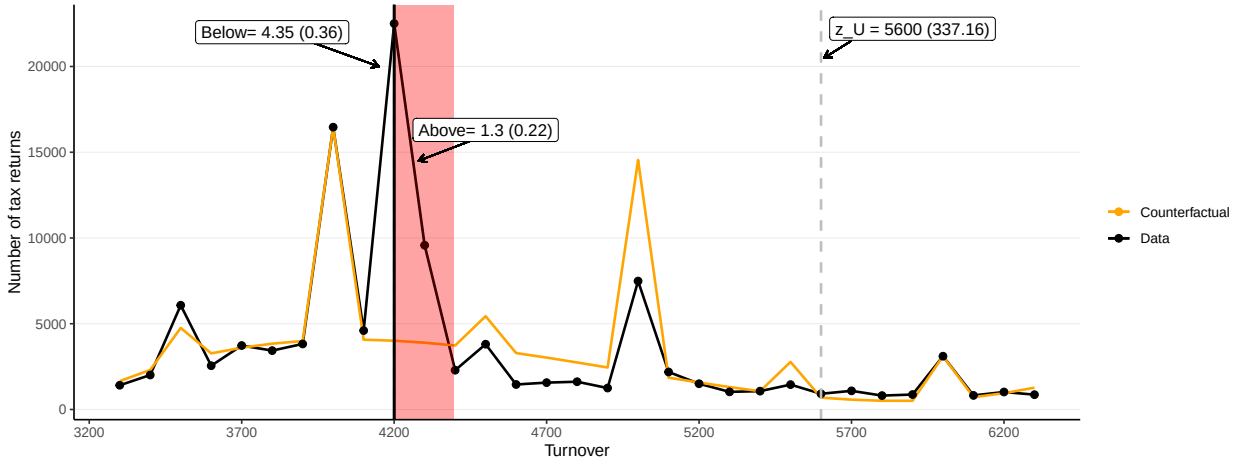
[8] If a firm, however, desires to pay 30 it would need to report then $\frac{30}{0.03} = 1000$. In this case, one could not know from the data whether the firm reported 1000 because of lazy reporting or because 30 was the desired payment amount.

in place. It is worthwhile to mention that the Zambian schedule during that time is particularly suitable to estimate behavioral responses to taxes as the thresholds at 4200, 8300, 12500, 16500 and 20800 neither represent very salient numbers of turnover, nor are they associated with any other kinds of discontinuities in terms of benefits or taxes. Methodologically, we estimate bunching around thresholds as explained. In particular, we compare the distribution of tax returns around the threshold to the estimated counterfactual distribution. The latter is derived from estimating Eq. (1) and then fitting the density $\hat{c}_j$ while omitting the effects of the γ_i's i.e. omitting the *affected range* around the notch. As illustrated in Figure 1, standard models of tax compliance predict the empirical density to exceed the counterfactual density below the notch (bunching) and to be substantially below the counterfactual above the threshold.

Figure 2 provides graphical evidence of the bunching patterns. It contrasts the empirical density with the counterfactual density across turnover bins of width 100 ZMK. The *dominated range* is marked as red. Clearly, the counterfactual density approximates the empirical density well up to the 4100 bin. Furthermore, beyond the estimated upper bound of the affected region z_U, the counterfactual approximates the empirical density well. Focusing on the area just below the notch, in bin $(4100, 4200]$, there is a strong spike in the empirical density exceeding the counterfactual density by far. The extent of bunching is estimated by Eq. (2) which gives a bunching coefficient of $B_{below} = 4.35$, implying that there are more than 5 times more tax returns at the threshold than is predicted by the counterfactual. The bootstrapped standard error of 0.36 shows its statistical significance at the 99% level. Appendix A shows that neither after nor before the notched schedule was in place, the distribution of tax returns featured bunching at these amounts.

When focusing on the region above the threshold, Figure 2 reveals another striking feature, which is at odds with the elaborated expectations. In particular, the empirical density also exceeds the counterfactual density in the first bin above the threshold. We quantify the extent of bunching *above* the threshold by following Eq. (3). The bunching estimate is $B_{above} = 1.3$. It is statistically significant at the 99% level with a bootstrapped standard error of 0.22. This bunching result is robust to using bins of different sizes as well as to relying on a completely non-parametric approach using the pre-reform distribution as the counterfactual (see Appendix A). Recall, that the standard theory, as delineated in Figure 1 predicts no tax return to appear within the dominated range, because the tax schedule offers clear incentives to reduce turnover to below the notch. Instead of a sharp decline above the notch in the density above the notch, we find strong and significant bunching within this

Figure 2: Bunching below and *above* the threshold



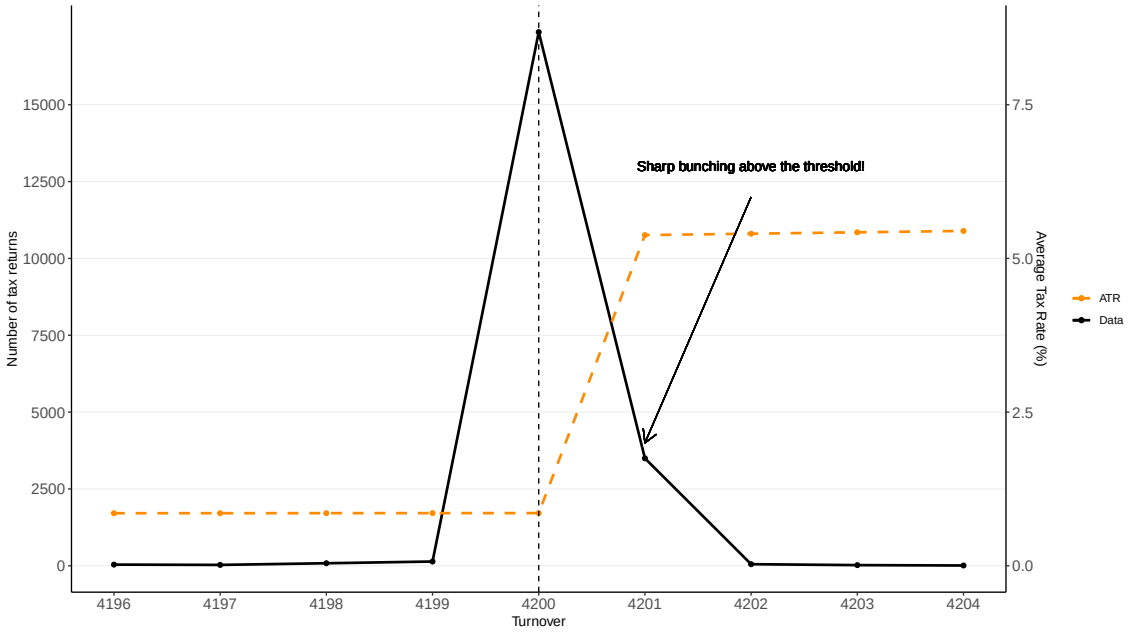
Notes: This figure plots the results of estimating bunching at the 4200 threshold. The black line depicts the empirical density. The yellow line depicts the counterfactual density as estimated by Eq. (1), accounting for round turnover amounts as well as round payment amounts. The black solid vertical line marks the threshold at which tax liability increases discretely i.e. the notch. The red area depicts the dominated range. The grey dashed vertical line depicts the upper bound of the omitted region z_U. Estimates of bunching below and above the threshold are derived from Eq. (2) & (3) and compare the counterfactual to the empirical density. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. Data source: ZRA. Years: 2017,2018.

range.

To shed more light on this unexpected bunching response, Figure 3 zooms in on the area around the threshold and plots the density of tax returns in substantially smaller bins of size 1 ZMK instead of 100 ZMK. The figure clearly shows that the bunching response above the threshold is very sharp. While there is a strong accumulation of tax returns between 4200 and 4201, the number of tax returns drops for turnover above 4201. To contrast, this with the incentives the tax schedule is offering at the notch, we plot the average tax rate. This underlines the puzzling bunching pattern: firms are declaring exactly 1 ZMK more, even though the average tax jumps from about 1% at 4200 to more than 5% at 4201.

In addition to the threshold at 4200, we estimate bunching responses at the other thresholds. Table 2 reports B_{below} and B_{above} for the all thresholds. The pattern of significant bunching below as well as *above* holds across all thresholds. Additionally, also the concentration of tax returns which report turnover sharply above the threshold is consistent throughout. In Appendix A, we provide the plots for these other thresholds and show that the extent of bunching patterns is very persistent over all 24 months in which the notched schedule was in place. The last row of Table 2 shows that a substantial fraction of tax returns within the first bins above thresholds, reported a turnover between the threshold and the threshold +1.

Figure 3: Sharp bunching *above* the threshold



Notes: The black solid line plots the empirical distribution of tax returns around the 4200 threshold with bin size of ZMK 1. The dashed orange line plots the average tax rate that firms are facing given their turnover. Data Source: ZRA. Years: 2017,2018.

To the best of our knowledge, we are the first to document bunching of tax returns in strictly dominated regions. It is important to note, however, that the literature has highlighted the role of optimization frictions which may lead firms to

Table 2: Bunching estimates

	Thresholds $\bar{k}$				
	4200	8300	12500	16500	20800
$\mathbf{B_{below}}$	**4.35**	**2.6**	**0.94**	**2.04**	**1.53**
	(0.36)	(0.33)	(0.21)	(0.43)	(0.13)
$\mathbf{B_{above}}$	**1.3**	**1.31**	**0.49**	**0.95**	**0.55**
	(0.22)	(0.21)	(0.17)	(0.42)	(0.08)
Sharp bunchers	**37%**	**37%**	**19 %**	**31 %**	**17%**

Notes: This table shows the estimated bunching coefficients for bunching above (Eq. (2)) and bunching *above* (Eq. (3)) for all notches. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. The last row depicts the share of sharp bunchers i.e. tax returns with turnover within the first bin above the threshold which report between the threshold and the threshold +1 i.e. bunch sharply above the threshold. Data source: ZRA. Years: 2017,2018.

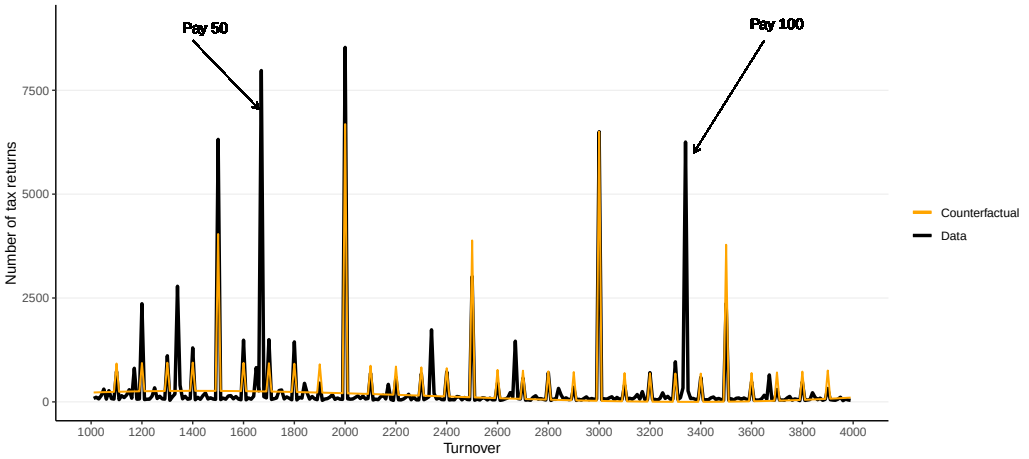
end up in dominated regions above notches. The idea is that frictions render firms unable to bunch below a notch exactly and thus they end up above the notch. However, there is no reason to expect that frictions will lead to bunching at a specific point in the dominated region. This is also confirmed by existing empirical evidence, which shows that frictions lead to diffuse mass and not bunching above thresholds (e.g., Kleven and Waseem, 2013; Anagol et al., 2022). We thus rule out optimization frictions as the main driver of these bunching results. Further evidence and a discussion on the role of optimization frictions are provided in Section 6.2.

Bunching at round number tax liabilities. We also estimate bunching at *round number tax liabilities*. To do so we, we pool all tax returns filed between 2015 and 2016 and proceed as described in Section 3.2. During these two years, the tax schedule featured a flat rate of 3% for all businesses irrespective of their size. This allows us to cleanly distinguish between bunching at round numbers of the tax base and bunching at round number tax liabilities. Figure 4 contrasts the the density of tax returns as observed in the data with the counterfactual density over the binned turnover distribution. The counterfactual density resembles the fitted values $\hat{c}_j$ as estimated by Eq. (4) without the influence of the round number tax liabilities (i.e. omitting ξ_n) but accounting for potential bunching at round number turnover. Bunching at round number tax liabilities is present whenever the density in data exceeds the counterfactual density in a given bin.

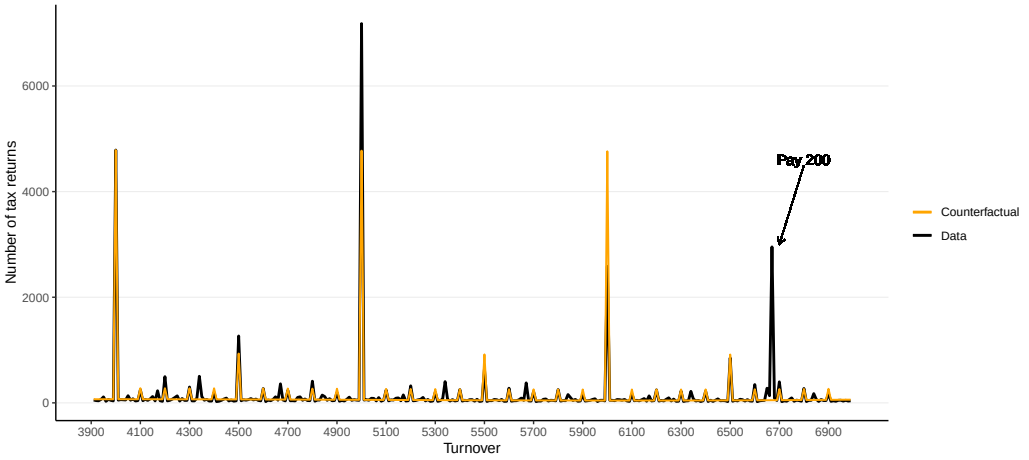
Panels (a) to (c) of Figure 4 provide striking visual evidence of such bunching. The empirical density features several mass points at which it exceeds the counterfactual density by far. Further, the extent of the bunching increases with the salience of "roundness". For example, while there is a strong bunching of tax returns at 1333.33 ZMK, which implies a liability of 40 ZMK (and is not captured by the counterfactual), the extent of bunching at 1666.66 ZMK, which implies a liability of 50 ZMK is nearly four times as large. The same observation holds for other multiples of 50, as highlighted in the figures. In most cases, our estimated counterfactual captures bunching at round number turnovers well and the discrepancies between empirical and counterfactual densities are small. In a number of cases, however, round turnover amounts and round number tax liabilites coincide. This is true for e.g. all multiples of 1000. Overall, bunching at round number tax liabilities is strong across a wide range of turnover. In Appendix B, we provide the corresponding bunching estimates for all liabilities which are divisible by 10, show that the patterns are persistent across time and further, that out of all tax returns filed between 2015 and 2016, 40% imply a tax liability divisible by 10.

Figure 4: Round Liability Bunching

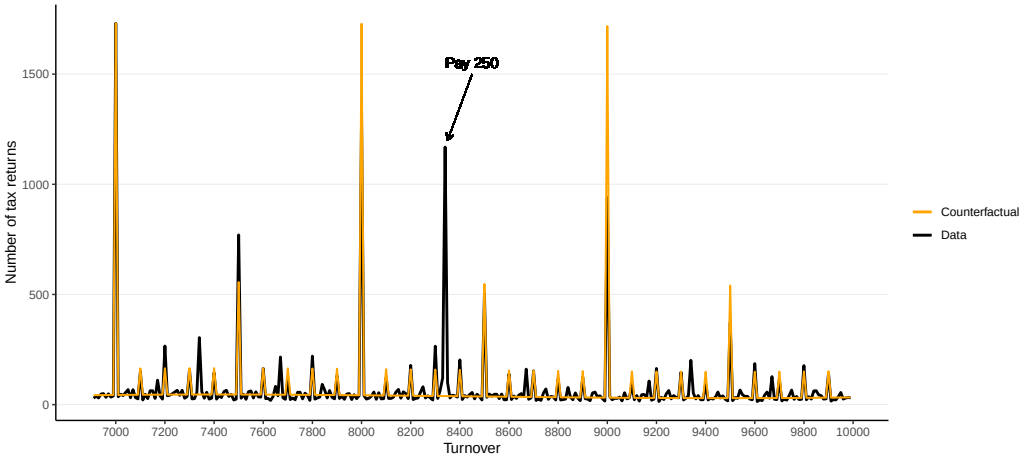
(a) Turnover 1000-4000



(b) Turnover 4000-7000



(c) Turnover 7000-10000



Notes: This figure plots the results of estimating bunching at round liability figures over the range 1000 to 10,000. The black line depicts the empirical density. The yellow line depicts the counterfactual density as estimated by Eq. (4). Bunching at payments of 50, 100, 200 and 250 are highlighted by arrows. Data source: ZRA. Years: 2015,2016.

To the best of our knowledge, we are the first to show clear evidence of bunching at round number tax liabilities. This type of bunching is yet undocumented in the literature, but highlights that the payment amount itself rather than the declared tax base is the salient amount when tax returns are filed.

Bunching at round liabilities predicts bunching above thresholds. In a next step, we show that the behaviors underlying these two empirical facts are linked. We start by identifying firms that ever bunch *above* a threshold in the years 2017-2018. Exploiting the panel structure of the data, we then regress an indicator for whether a firm has ever bunched *above* on the share of the firm's tax returns prior to 2017 that were rounded at the liability level, i.e. imply a liability divisible by 10. To clearly distinguish from rounding at the turnover level, we do not include returns with a turnover divisible 1000, which implies both round liability and round turnover.

We include control variables across multiple dimensions.[9] First, we include several other pre-reform characteristics. Analogous to the variable capturing rounding at the liability level, we calculate the share of tax returns with round turnover amounts prior to the reform. We further calculate the share of tax returns with turnover that is exactly equal to the firm's previous month's tax return. This variable captures the phenomenon that firms in lower income countries may rely on past tax returns as a heuristic and simply file the same amount again (cf. Tourek (2022)). Finally, we account for the extent of pre-reform compliance (i.e. how often a tax return has been filed). Second, we make use of the audit data to control for whether a firm has ever been audited audited and whether a payment of penalties or owed tax had to be made. We include sector- and tax office fixed effects for all regressions.

Table 3 presents the results. The different columns correspond to 3 different specifications, gradually including more controls. In all specifications, the strongest and only statistically significant predictor for whether a firm ever bunches above is exerting round tax liabilities. Overall, the results imply that a 1 percentage point (pp) increase in the share of tax returns with a round liability is associated with a 0.3 pp larger probability of bunching above a threshold. This implies an increase of 16% relative to the baseline. This effect is stable and significant at the 99% level throughout all three specifications. Further insights can be gained by inspecting the effects of the other regressors. First, we observe no significant effect of turnover

[9] As shown in Appendix A, neither any specific sector nor any specific geographic location are significant predictors. Regressions on either tax-office or –sector fixed effects deliver a R^2 of about 0.001.

Table 3: Predictors for bunching *above*

	Dependent variable:		
	Ever bunched above (0/1)		
Firm characteristics	(1)	(2)	(3)
Liability rounding	**0.165*****	**0.168*****	**0.167*****
	(0.011)	(0.012)	(0.012)
[Liability × Turnover] rounding	**0.147*****	**0.152*****	**0.151*****
	(0.011)	(0.012)	(0.012)
Turnover rounding	**−0.006**	**−0.003**	**−0.004**
	(0.011)	(0.011)	(0.012)
Targeting past payments		**−0.014**	**−0.013**
		(0.016)	(0.016)
Audited (0/1)			**0.015**
			(0.056)
[Audited × Penalty] (0/1)			**0.097**
			(0.064)
Pre-reform compliance			**0.00003**
			(0.0003)
# Firms	18,516	18,516	18,516
Taxoffice FE	✓	✓	✓
Sector FE	✓	✓	✓
Baseline mean	**0.019**	**0.019**	**0.019**
R^2	0.023	0.023	0.023

Notes: This table shows the estimated correlations between bunching *above* thresholds and firms' behavioral characteristics prior to the notched schedule (Years 2015, 2016). 'Liability rounding' denotes the share of a firm's tax returns which are rounded at the liability level (with a round liability but odd turnover, e.g. 3333.33). 'Turnover rounding' denotes the share of tax returns rounded at the turnover level. 'Targeting past payments' is the share of tax returns with turnover that is exactly equal to the firm's previous month's tax return.'Audited' is an indicator for whether a firm has been audited. 'Penalty' is an indicator for whether a firm had to pay a penalty due to tax reasons. 'Pre-reform compliance' measures how often the firm has filed taxes prior to the reform. Data source: ZRA. Years: 2015-2018.

rounding, speaking against the explanation that the observed phenomenon is a mere issue of bad record keeping. Second, targeting past payments is not a significant predictor of bunching above suggesting that the heuristics on past payments plays a minor role. Third, the experience of being audited itself does not have a large influence on the outcome variable. Though, we do see a sizable effect of whether

a firm has had penalty payments from tax audits, this effect is insignificant[10] and importantly its inclusion in the regression does not alter the effect of the strongest predictor: liability rounding.

4.2 Results from a firm survey

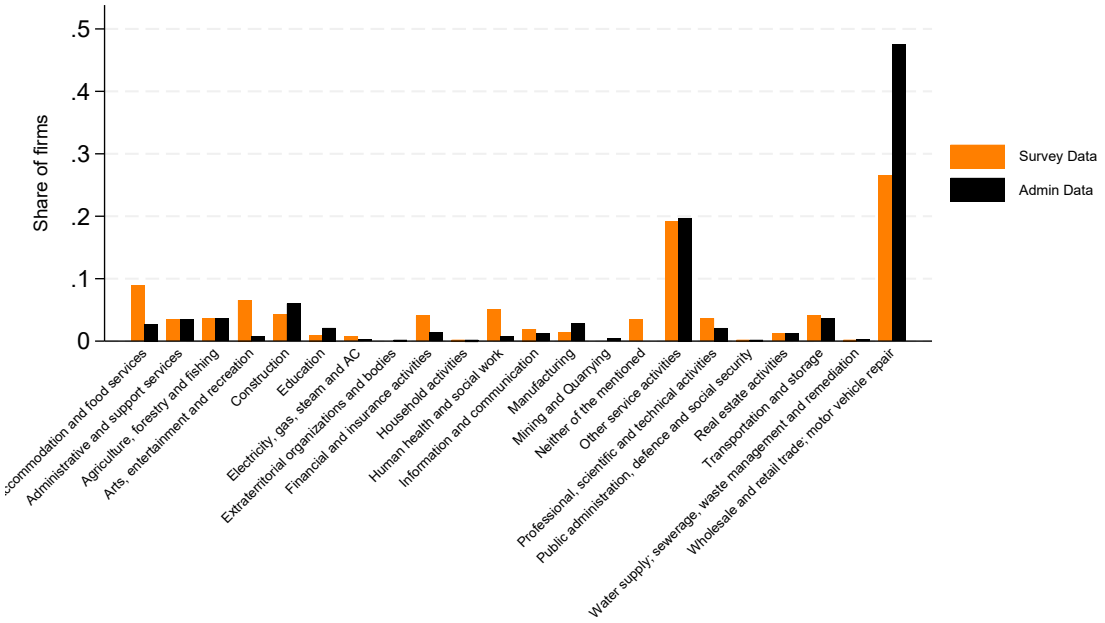
The empirical facts established from the administrative data show a stark contrast to the behaviour of firms under standard assumptions. However, as is typical for administrative data, it contains only limited information on firms' characteristics and no information on tax compliance behavior. Thus, we complement our analysis with findings from our own survey, which we described in Section 3.1.

Sample. Overall, 517 firms were surveyed in Lusaka in November and December 2023. Firm owners were quasi-randomly selected and approached in-person by the surveyors. The quasi-random selection followed a "snow-ball procedure". Once consent was given, the interview of approximately 30 questions started. A size restriction on a firm's turnover was put in place to ensure capturing only small firms. Figure 5 compares the distribution of sectors across firms in the survey with the administrative records. As the share of firms within most sectors aligns well, the survey can be considered representative for the firm population.

Firm characteristics. Panel A of Table 4 provides information on general firm characteristics. Half of all firms had not more than 2 employees (25% had no employees) and on average monthly turnover was ZMK 24,000. In 35% of cases, the owner was female and the most common sector was wholesale & trading. The results show that the vast majority of firms keeps detailed records of sales (i.e. turnover) and more than half of those keep track of those in a more proficient way than solely on paper logs. Interestingly, an even larger share states to keep detailed records of costs. As costs are irrelevant for turnover tax purposes but relevant for disposable income, this suggests that firms care more about what is paid in taxes ultimately rather than what the correct turnover is. It also suggests that firms are not generally lacking the resources to keep sales records. Further, tax literacy appears to be widespread. 96% state the correct turnover tax rate when asked and 80% are educated to at least a college degree. Finally, as opposed to a key assumption in standard models, 93% of surveyed firms state not to incorporate their tax liability

[10] Audits have very low explanatory power, even when taken at face-value. The R^2 when only keeping audit variables as regressors is 0.003.

Figure 5: Industry compositions of firm samples; by data source



Notes: This figure plots the distribution of firms across industries in both the survey data as well as the administrative data. Data: ZRA (administrative data) and own collected data (survey data). Years: 2015-2021 (administrative data) and 2023 (survey data).

when making business decisions. This implies that the marginal tax rate is irrelevant for the firm's optimization problem and thus, doing business and deciding on the reported tax base pose two separate problems.

Tax behaviour. Panel B of Table 4 concentrates on tax behavior. The most striking pattern which emerges is that discussions between officials and taxpayers about the tax payment is commonplace. Focusing on firms that were active during the period in which the notched schedule was in place (2017, 2018), 55% of respondents said they discuss the payment they are going to make with an official *before* filing their tax return or that such discussions are common. Overall, the share of firms stating to discuss payments with tax officials is 42%. Out of these responses, more than a third explicitly stated that during such discussions taxpayer and tax official "find an agreement" on what should be paid. In other words, firms are bargaining over tax payments with tax officials. As we show formally in Section 5, the prevalence of bargaining over tax payments can rationalize the seemingly irrational empirical facts in the administrative data. Interestingly, there is a strong and significant negative correlation between discussing tax payments and the perceived probability of being audited. This could hint at the fact that for the tax authority, bargaining

Table 4: Characteristics and tax behavior of surveyed firms

	Mean (1)	N (2)
Panel A: General		
Female (owner)	0.35	517
College or higher (owner)	0.8	517
# employees (median)	2	517
monthly turnover in (ZMK, average)	24,000	517
Firm does business with government	0.25	517
Whether firm keeps detailed records of sales	0.89	517
- only on paper	0.39	451
Whether firm keeps records on expenditures	0.95	517
Whether firm knows turnover tax rate	0.96	515
Whether firm thinks about tax liability when making business decisions	0.07	517
Panel B: Taxes		
Whether firms (including self) discuss payments with tax officials before filing		
All firms	0.42	517
- "to agree on payments"	0.36	216
- "to clarify tax liability"	0.57	216
Firms active during notched schedule (2017, 2018)	0.55	193
Perceived probability of being audited for tax reasons	0.19	515
	Coef.	p-value
- correlation with firm discussing payments	-0.05	0.01
Perceived percentage of firms bribing tax collectors	0.15	257

Notes: This table provides results from the survey conducted on firms under turnover tax in Lusaka between November and December 2023. Source: survey data.

over taxes before returns are being filed may be a substitute to formal audits after returns have been filed. Regarding corruption, we find that 15% of firms are deemed to pay bribes to tax collectors. This is much lower than the share of firms discussing payments with officials, suggesting that discussions do not serve a mere corruption purpose.

5 A model of tax bargaining

In this section, we show that the prevalence of bargaining rationalizes the observed bunching patterns. We begin by verbally delineating our main argument before formally introducing a theoretical model. Alternative explanations, which we rule out, are discussed in depth in Section 6.

5.1 Main idea

In many lower income countries, including Zambia, discussions with tax officials play a crucial role for tax compliance of small firms (cf. Section 2). Our survey has elicited that these discussions often serve the purpose of discussing intended tax payments before returns are filed and thus can be interpreted as bargaining situations. Against this background, it is clear that, what is observed as turnover in the tax data has a very different interpretation. In particular, rather than representing true turnover as a result of economic activity, tax returns indicate a turnover that results in a payment which is *agreed upon* by taxpayers and tax officials. We argue that following this interpretation is important to rationalize the observed patterns.

To see the argument, it is useful to consider bunching at round number tax liabilities first. This bunching pattern aligns well with the idea that tax returns represent the outcomes of bargaining. First of all, it highlights the focus on the actual payment as the salient amount when filing taxes. Further, a large literature has highlighted that round numbers serve as focal points in bargaining situations (e.g., Schelling, 1960; Albers and Albers, 1983; Janssen, 2001, 2006; Pope et al., 2015). Thus, the strong accumulation of tax returns at turnover amounts which imply round payments may indicate that these are bargained and agreed upon payments. For example, when bargaining, it is more likely that two parties would settle on a payment of e.g. 100 ZMK instead of 99 ZMK.

Against this background, we turn to rationalize the fact that firms bunch above notches – a strictly dominated dominated choice through the lenses of standard models. The analysis so far suggests that firms and tax officials bargain over tax payments and the declared turnover represents the *inverted* tax schedule to arrive at the agreed upon payment. Under a flat tax schedule any such payment can be reached. When the tax schedule features notches, however, these notches effectively introduce regions of payment that cannot be reached anymore. To see this, one can consider the example of the 4200 threshold in the Zambian case (cf. Table 1). When declaring 4200 ZMK as turnover, the payment will be 36 ZMK. But when declaring a bit more, e.g. 4201 ZMK, the payment will be 225 ZMK. Thus, it is not possible to implement a payment between 36 and 225. When firms and tax officials are bargaining and would have settled on a payment within this interval, the agreed upon payment will end up either below or above that interval. In the tax data, this will create bunching on both sides of the notch. In what follows, we provide a theoretical framework to formalize this idea and offer a microfoundation for bargaining.

5.2 Theoretical model

To study how and when bargaining over taxes between taxpayers- and collectors evolves, we begin by establishing equilibrium tax payments in a non-cooperative setting (without bargaining). Then, we proceed to characterize pareto-improving payments in a cooperative setting (with bargaining). We show how the bargained payments relate to true tax liability and how the scope for bargaining vanishes along the path of economic development. Finally, we demonstrate how bargaining rationalizes the empirical facts described in Section 4.1.

Environment and agents. We consider a taxpayer with turnover z. The taxpayer's disposable income is given by $y = z - T - \pi$, where T denotes the tax payment and π any potential penalties. We will refer to the firm owner as the taxpayer in the following. Her preferences over consumption are described by $U_F = v(y)$ with $v'() > 0$ and $v''() < 0$. We further assume that the inverse of both $v()$ as well as $v'()$ exist. Given turnover z, the taxpayer needs to decide on the tax payment T. We assume that the firm will never pay more than it legally owes and thus $T \leq T(z)$. If T deviates from the the true liability $T(z)$, such that $T < T(z)$, the taxpayer risks, that, when being audited, the tax authority notices the discrepancy. In that case, she has to pay a fine (then $\pi > 0$). We denote the joint probability of being audited and caught by p. The fine that has to be paid if caught is proportional to the evaded amount by factor ξ (cf. Allingham and Sandmo (1972)). The taxpayer's maximization problem thus reads:

$$\max_{T} \quad \mathbf{E}[U_F] = (1-p)v(z-T) + pv(z-T-(T(z)-T)(1+\xi)) \tag{6}$$

Denoting the solution to (6) as T^*, we can write the expected tax payments and fines in the *non-cooperative* setting as

$$T^{nc} \equiv (1-p)T^* + p(T^* + (T(z)-T^*)(1+\xi)). \tag{7}$$

On the receiving end of the tax payment is the tax authority. We assume that it has linear utility U_G in tax payments and fines and costs of engaging in audits. The latter is described as $c(p)$ with $c'() > 0$ and $c''() > 0$, accomodating the notion that increasing the probability of detecting tax evasion is increasingly costly. Let κ be the elasticity of costs with respect to the implemented audit and detection probability

p: $\kappa = \frac{\partial c(p)}{\partial p}\frac{p}{c(p)}$. We assume that $\kappa \geq 1$.[11] The decision problem of the tax authority reads:

$$\max_{p} \quad \mathbf{E}[U_G] = T^{nc}(p) - c(p). \tag{8}$$

Figure 6 depicts a numerical example of both utility functions graphically as a function of the tax payment. The expected utilities in the non-cooperative equilibrium are denoted by $E\left[U_F|T^{nc}\right]$ and $E\left[U_G|T^{nc}\right]$ respectively.

Pareto-improvements through bargaining. We now consider the option of taxpayer and –authority bargaining over tax payments. As shown in our survey results (cf. Section 4.2), such behaviour is a common feature of how tax payments are determined. We thus extend the model as follows. If both taxpayer and tax authority can agree on a tax payment in advance, then no audits will take place. Naturally, an agreement may only be reached if both parties are better off than without bargaining. If there is no agreement, the outside option realizes which is the non-cooperative equilibrium as described above. We therefore consider the set of payments which are pareto-improving relative to the non-cooperative equilibrium. Clearly, the taxpayer would agree on any payment T such that $v(z-T) \geq E[U_F|T^{nc}]$ i.e. she is as least as well off as in the non-cooperative setting. We define the maximum amount of certain payment the taxpayer would be willing to make as T_F, derived from the following inequality:

$$v(z-T) \geq E[U_F|T^{nc}] \Longleftrightarrow T \leq z - v^{-1}(E[U_F|T^{nc}]) \equiv \mathbf{T_F}. \tag{9}$$

For the tax authority, agreeing on a certain tax payment saves the audit costs $c(p)$. On the other hand, costs related to bargaining with the taxpayer may occur. We denote the costs of audits net of bargaining costs as $\tilde{c}(p)$.[12] Note that if $\tilde{c}(p) > 0$, bargaining shifts the utility function of the tax authority upwards as depicted in panel (b) of Figure 6. Thus, the minimum tax payment the tax authority would accept is given by T_G which can be derived from the following inequality:

$$T \geq T^{nc} - \tilde{c}(p) \equiv \mathbf{T_G}. \tag{10}$$

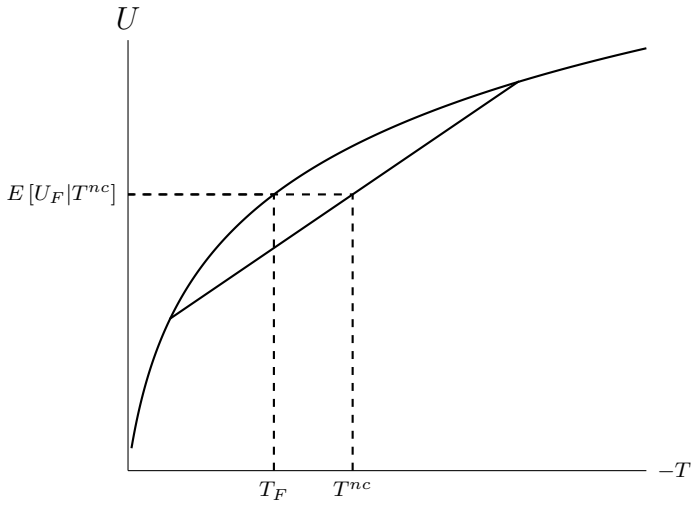
Both T_F and T_G are depicted in Figure 6.

[11] Technically, this implies that increasing the audit probability by 1%, increases the cost by at least 1%.

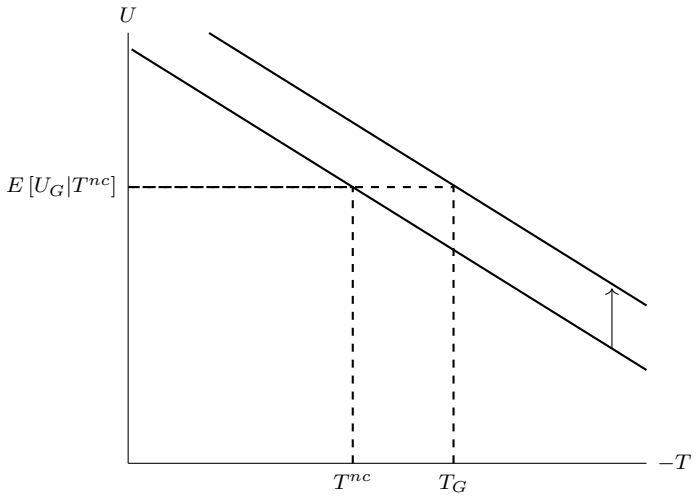
[12] For simplicity, we set bargaining costs for firms to 0. However, this assumption is not essential to our results.

Figure 6: Utility functions and outside options in the bargaining model

(a) Taxpayer



(b) Tax Authority



Notes: This figure illustrates the utilities of taxpayer and tax authority as a function of the tax payment T. T^{nc} denotes the equilibrium payment in the setting without bargaining. T_F and T_G depict the outside option payments and thus determine the set of pareto-improving payments.

Equilibrium. An equilibrium with bargaining can only exist if taxpayer and tax authority are at least as well off as in the non-cooperative setting. Thus it needs to hold that:

$$T_F \geq T_G. \tag{11}$$

Intuitively, this means that bargaining is only beneficial if the maximum amount the taxpayer is willing to pay is larger than the minimum amount the tax authority is willing to accept. The bargaining set is therefore given by the interval (T_G, T_F). In Proposition 1, we formalize how the true tax liability $T(z)$ relates to the set of

possible bargaining outcomes.

Proposition 1. *The true tax liability $T(z)$ is outside the bargaining set. In particular $T(z) \geq T_F$.*
Proof: See Appendix F.

Two insights can be derived from Proposition 1. First, taxpayers engaging in bargaining, can reduce their de-facto tax payment without inducing the risk of penalties through audits. Second, the payment resulting from bargaining is detached from a taxpayer's true liability, which is consistent with the notion that bargained payments often simply end up on round figures.

We turn to ask under which circumstances the bargaining set is non-empty and scope for bargaining exists. In particular, why we would not expect such bargaining to happen in more developed countries. One key feature of our model that may change along the path of development is the costliness of doing successful audits $c(p)$. In fact, the turnover tax for small firms outside the VAT net is difficult to assess and therefore to audit. This is because most transactions are made in cash and leave no paper trail. There is therefore virtually no third party reporting in most cases. In more developed countries, third-party reporting and withholding is more common and successfully assessing and auditing a large share of a firm's income is easy for tax authorities (Kleven et al., 2016). In our model, the costliness of doing successful audits can be characterized as the elasticity of the cost function with respect to audits. A lower κ is therefore associated with larger state capacity. The role of κ for the bargaining outcome is summarized in Proposition 2.

Proposition 2. *With $\kappa \to 1$, the bargaining set collapses and $T_F = T_G = T(z)$.*
Proof: See Appendix F.

The intuition behind Proposition 2 is that as countries develop and their state capacity grows, they are able to enforce taxes more efficiently. In its limit, our model predicts that the set of possible bargaining outcomes breaks down.

Dividing the surplus. If condition (11) holds, then there exists a surplus from bargaining which is given by $T_F - T_G$. We assume that this surplus is divided

among taxpayer and tax authority via Nash-bargaining.[13] Clearly, the most desirable outcome for the taxpayer would be to pay only T_G, i.e. the minimum amount the tax authority would accept. The most desirable outcome for the tax authority is the extract the maximum amount the taxpayer is willing to pay, which is T_F. Denoting the bargaining power (this can be thought of e.g. as enforcement capacity) of the tax authority as α, the following Nash-product is maximized:

$$\max_T (T_F - T)^{1-\alpha}(T - T_G)^{\alpha}. \tag{12}$$

Rewriting the first-order condition of Eq. (12), we arrive at the *cooperative* solution

$$T^c = \alpha T_F + (1-\alpha)T_G.$$

Bunching in the bargaining model. We now turn to illustrate how the bargaining framework can rationalize the empirical bunching patterns. First, we consider the linear tax rate setting: Once the firm and the tax authority have agreed on a payment $x = T^c$, the firm needs to report $T^{-1}_{schedule}(x)$. As the tax rate is at 3%, this is simply $\frac{x}{0.03}$. A linear tax schedule is bijective and therefore, every payment that was agreed upon could be reached. The behavorial literature has shown that bargaining outcomes are strongly concentrated on round figures, supporting the notion that bunching at round tax payments stems from bargaining (Schelling, 1960; Pope et al., 2015; Albers and Albers, 1983).

Next, we consider the notched tax schedule (cf. Table 1). As notches imply jumps in the average tax rates, they effectively introduce regions of tax liability that can not be implemented. We illustrate this at the example of the 4200 threshold:

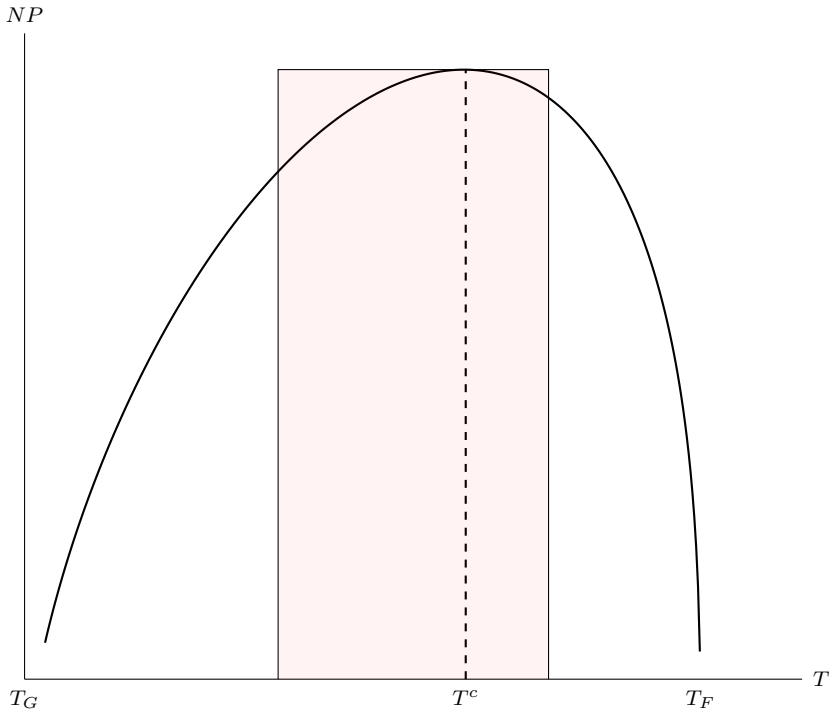
$$T_{schedule}(4200) = 36;\ T_{schedule}(4200.01) = 225$$
$$\Longleftrightarrow$$
$$T^{-1}_{schedule}(x) = \emptyset,\ x \in (36, 225)$$

Now, for all firms with $T^c \in (36, 225)$ the bargaining problem becomes a case distinction of whether the Nash product in Eq. (12) is larger above or below the threshold.[14]

[13] Nash-bargaining offers a convenient way to ensure uniqueness of the bargaining outcome. Other bargaining models incorporating e.g. focal points (Janssen, 2001) may be considered in future work.

[14] Eq. (12) is strictly concave in T under the assumption that $T_F \geq T_G$. Therefore the maximization problem turns into a case distinction between the two options if the unconstrained optimum is within $(36, 225)$.

Figure 7: Nash-Product with notch



Notes: This figure illustrates the Nash-product (NP), which the tax authority and taxpayer maximize when bargaining over the payment (concave curve). T^c denotes the unconstrained solution. The red rectangle illustrates a region of payments which become unreachable due to a notch in the tax schedule. In the illustrated case, the NP is larger *above* the notch than below. Therefore, the declared turnover needs to be just above the threshold.

Thus, firms within this range report turnover either above or below the threshold. Tax returns will accumulate on both sides which generates the observed bunching. If we consider T_F and T_G as functions of true turnover z, the cutoff for bunching *above* is at $\tilde{z}$ such that

$$(T_F(\tilde{z}) - 36)^{(1-\alpha)}(36 - T_G(\tilde{z}))^{\alpha} = (T_F(\tilde{z}) - 225)^{(1-\alpha)}(225 - T_G(\tilde{z}))^{\alpha}.$$

The case of Nash-bargaining over tax payments in the presence of notch is depicted in Figure 7. In this graphical example, the notch leads to the bunching *above* the threshold.

Finally, let true turnover z be continuously distributed with density f. The total bunching mass on both sides of cutoff is given by

$$\underbrace{\int_{z_B}^{\tilde{z}} f(z)dz}_{\textbf{Bunching below}} + \underbrace{\int_{\tilde{z}}^{z_A} f(z)dz}_{\textbf{Bunching above}} \tag{13}$$

where $\tilde{z}$ is the cutoff value from above, z_A is such that $T^c(z_A) = 225$ and z_B such

that $T^c(z_B) = 36$. Eq. (13) therefore provides a mapping of the distribution of true turnover to the bunching we see in the data.

Insights. Our model shows that bargaining over taxes is pareto-improving over the non-cooperative standard setting under reasonable assumptions for lower income countries. This holds because firms are risk averse and/or the tax authority has higher costs of auditing ex-post than discussing payments with firms ex-ante. Note that each condition implies room for bargaining independently of each other. We demonstrated how the model rationalizes the two empirical facts, which are otherwise puzzling. First, bargained payments are detached from a firm's true liability, corroborating the notion that payments end up on round numbers. Second, notches restrict the feasible bargaining set, leading to payments bunching above the threshold.

5.3 Discussion

We will now discuss the role of corruption and the incentives of tax collectors in the process of bargaining.

The role of corruption. Interactions between taxpayers and tax officials are inherently connected to concerns about corruption. One might worry that a system, which allows for such interactions supports corruption, where firms can bribe tax officials and make lower tax payments. While we cannot rule out that such activities are taking place, we argue that corruption is unlikely to be the main driver. First, from our survey as well as the World Bank Enterprise Surveys, we see that meetings and discussions with tax officials are much more common than bribery of tax officials. The World Bank estimates that in Zambia, more than 80% of small firms regularly visit tax officials, but less than 4% bribe tax officials during such visits (cf. sections 2 and 4.2). This suggests that the discussions generally serve another purpose than exchanging bribes. Bargaining and eventually agreeing on a tax payment as outlined above is one such purpose, supported by the results of our own survey. One should also note, that bunching above a threshold is hardly consistent with corruption. If firms and tax officials would collude, they would always have the clear incentive to agree on a tax payment below the threshold and share the difference to the payment that would have been due above the threshold. Hindriks et al. (1999) provide a theoretical framework for corruption which formalizes this

notion.

Incentives of tax officials. If corruption does not play a major role, one could ask, what the incentives of tax officials are besides that. First, it is to note that bribery is obviously a misconduct and tax officials could lose their (relatively well paid) job.[15] Second, the ZRA has a system of performance benefits in place which rewards tax offices with good collection performance. These benefits, however, can only be paid out to individual tax officials when the tax office reaches its prescribed revenue target. Thus, thanks to the incentive scheme, incentives of officials and tax authority are aligned. In a bargaining situation, this incentive could lead tax officials to push for firms to bunch above the threshold. This would further mean that once the revenue target is reached, one would observe less bunching above. We confirm this channel empirically and estimate a significant drop in the likelihood of firms bunching above once a tax office has hit its revenue target (cf. Appendix D).

6 Alternative explanations

This section points out several alternative explanations for the empirical facts and shows that each of them are unlikely to be sufficient explanations.

6.1 Audits

One alternative explanation might be that even without bargaining, the threat of being audited might motivate firms to pay more and therefore bunch above the threshold. We provide four pieces of evidence to rebut this concern.

Correlational Analysis. As seen in Table 3, we do not find a significant relationship between a firm ever being audited and a firm ever bunching above. While there is indeed a sizable point estimate for whether a firm has ever paid a fine, it is imprecisely estimated. Note that even if we would take this estimate at face-value it does not explain a lot of the variation in the explanatory variable. This is because the number of firms that ever bunch above by far exceeds the number of firms that have ever paid a fine after being audited.

[15] The median annual salary for tax collectors in Zambia is more than twice as much as the country's GDP per capita. Source: `https://worldsalaries.com/average-tax-officer-salary-in-zambia/`. Last accessed: June 21st, 2024.

Empirical audit probabilities. If firms would strategically bunch above thresholds to avoid audits, then the probability of being audited should be substantially lower above than below the threshold for it to be worthwhile to make higher tax payments. In Appendix C, we estimate empirical audit probabilites conditional on reported turnover and show that they do not differ substantially above vs. below a threshold.

Simulations from tax evasion model. We further address the concern that these empirical audit probabilities might not be known to taxpayers and differences might be perceived as larger. In Appendix C, we simulate a model of tax evasion along the lines of Kleven et al. (2011) with endogenous audit probability for the case of the Zambian turnover tax. We demonstrate that even under extreme assumptions where the perceived probability of being audited jumps from 10% above to 30% below the threshold, only firms that evade at least 83% of their turnover would choose to stay above the threshold.[16]

Randomized survey experiment. Finally, we took an experimental approach to test whether perceived audit probabilities could drive firms to bunch above thresholds. At the end of our survey, we included a randomized experiment with the aim to test the audit channel and also a government contracting channel. The treatment consisted of providing information and thereby shifting beliefs of respondents.

The experimental setup was as follows. We randomly assigned each respondent to one of three groups: control group or one of two treatment groups. We thereby followed the standard methodology of information experiments. First, it was tested whether the threat of audits might motivate firms to bunch *above* the threshold instead of below. For example, if audit probability increases discretely when moving below the threshold, firms might trade-off higher tax payments with lower audit probability below the threshold. The first treatment group therefore received information on the amount of audits and money seized therein by the ZRA. Second, as 75% of respondents stated that securing a business contract with the government was very competitive, we investigated whether this could be driving firms to put themselves in a good position with the government and bunch *above* the threshold rather than below. The second treatment group therefore received information on the amount of business the government is doing with SMEs and a reminder that one

[16]This result is for the standard model which assumes quasi-linear utility (Kleven et al., 2011; Kleven and Waseem, 2013; Bachas and Soto, 2021). Accounting for risk aversion, we conclude that only firms evading at least 58% or 50% of their turnover would choose to bunch above, when imposing CRRA utility and a risk-aversion parameter of 2 or 3, respectively.

needs a tax clearance certificate to engage in such contracts. The control group only received information about the total number of firms registered under turnover tax.

In order to measure a firm's propensity to bunch above the threshold, the survey followed a specific procedure. In the beginning of the survey, the average turnover of the firm was inquired. Later, the firm was asked which tax payment the respondent thinks is appropriate for a firm with the previously stated turnover. Then the randomized information treatment occurred. Once the information was given, respondents were confronted with a hypothetical notch: they faced a situation in which their stated appropriate payment is not feasible anymore, instead they only had the option to either pay 10% more or 10% less. Opting for the larger payment would then be associated with bunching above.[17]

We estimate treatment effects of of our two information treatments by running the following regression.

$$y_i = \alpha + \beta_1 treat_1 + \beta_2 treat_2 + \mathbf{X}_i + \gamma_e + \epsilon_i \tag{14}$$

where y_i denotes the response of respondent i after being treated, $\mathbf{X_i}$ are controls, γ_e are enumerator fixed effects and ϵ denotes the error term. $treat_1$ and $treat_2$ correspond to indicator for whether the respondent belongs to treatment group 1 or 2 respectively. We then test the null hypothesis $H_0 : \beta_1 = 0$ or $H_0 : \beta_2 = 0$ to investigate whether the treatments have a causal effect on the response y.

Table 5 shows the treatment effects of both information messages as estimated from Eq. (14). The last two columns of the table show results for specification where the the average turnover is fixed at 10,000 and the appropriate payment is exogenously shifted to 500 (column (4)) such that both feasible payments (450 or 550) are legal under the contemporary tax rate of 4% as well as to 300 (column (3)) where both options would be illegal. This serves as a check that responses are not driven by such legal considerations.

The results are striking in the sense that we estimate no significant effect of any treatment throughout. While the baseline shows no significant effect, also adding controls for turnover does not change this result in column (2). Further, neither columns (3) or (4) show a statistically significant impact on of the treatment on propensity to bunch *above*, suggesting that the result in columns (1) and (2) is not driven by legal considerations. Note that our confidence intervals also exclude

[17] As we are unable to link respondents' identities to actual tax records, we had to rely on this stated measure of the propensity to bunch above threshold. A more detailed description of the survey and examples are provided in Appendix E.

Table 5: Informational treatment effects on *bunching above*

	Dependent variable				
	Bunch above threshold				
	Baseline Liability:				
	Stated			ZMK 300	ZMK 500
	(1)	(2)	(3)	(4)	(5)
Audit Treatment	**0.0261**	**0.0287**	**0.0263**	**0.0124**	**0.0198**
	(0.0299)	(0.0304)	(0.0305)	(0.0325)	(0.0446)
Contract Treatment	**0.0110**	**0.0158**	**0.0149**	**0.0140**	**0.00952**
	(0.0294)	(0.0301)	(0.0297)	(0.0312)	(0.0426)
# Firms	517	510	510	251	259
Size control		✓	✓	✓	✓
Enumerator FE			✓	✓	✓

Notes: This table shows the results from the randomized survey experiment. The coefficients in columns (1), (2), and (3) represent the estimated effect of the two information treatments on an indicator for whether survey respondents would choose a larger tax payment if their favored option was no longer feasible. We interpret this as the propensity to bunch above. Additionally, column (2) controls for turnover. Columns (4) and (5) represent the same coefficients only that the initial tax payment option was not chosen by the respondents but fixed at ZMK 300 or ZMK 500 respectively. Robust standard errors are in parentheses. Data source: survey data. Year: 2023.

existing estimates on the effect of deterrence messages on tax compliance.[18] Overall, this null-result rejects the hypothesis that standard explanations such as the threat of being audited can rationalize the empirical facts.

Clearly, as in any survey experiment, the outcomes we measure are only stated and do not necessarily coincide with real actions. However, being unable to match survey respondents to their administrative tax records, we consider our approach to be second-best.

6.2 Frictions

We will now discuss the role of frictions which may hinder firms to respond to incentives optimally. This could lead firms that actually wanted to bunch below to bunch above a threshold, instead.

[18] A comparison to existing estimates of the effect of deterrence messages (such as increasing the salience of audit probabilities) is not straightforward for two reasons. First, the measure of increasing salience is heterogeneous across studies. Second, most studies measure effects on taxes paid. We are interested in bunching above thresholds. To still offer a comparison, we can translate the outcome variable of *bunching above* into an equivalent increase of tax paid by 10%. A positive effect of e.g. 5.6% (the upper bound on our confidence interval in column 1, row 1) would thus imply an increase in taxes paid by 0.56%. This is substantially lower than existing estimates from lower income countries (e.g., Shimeles et al., 2017; Mascagni and Nell, 2022) and also richer countries (e.g., Slemrod et al., 2001; Fellner et al., 2013)

Optimization frictions. Potentially, taxpayers locating above the threshold, were actually aiming to have a turnover just below the threshold but simply failed to do so. This idea has been formalized in the literature already and in the following, we will consider two such approaches and show that optimization frictions are unlikely to drive the observed bunching pattern.

Kleven and Waseem (2013) explain the presence of mass in the dominated region by optimization frictions. In their application, sharp bunching *below* the threshold suggests an extensive margin of frictions: either a taxpayer is able to manipulate turnover to lie below the threshold (in that case, exactly) or the taxpayer is unable to adjust at all. In the case of Zambia, the bunching *above* the threshold is also sharp. Bunching above therefore seems to be a strategic choice rather than driven by frictions to bunch below the threshold.

A more recent paper by Anagol et al. (2022) explains excess mass in dominated regions by the probability distribution of opportunities around the threshold. The idea is that given the threshold as a target, taxpayers draw from a discrete set of opportunities around the target. This might result in some taxpayers "overshooting" the target and ending up above the notch. However, as in Kleven and Waseem (2013), this model also predicts a hole in the distribution above the notch, rendering it clearly inconsistent with the excess mass in this area. We conclude that neither type of optimization frictions discussed in the literature is able to explain the observed bunching pattern.

Record keeping. Another form of frictions could be a firm's low capacity to keep detailed records of turnover and therefore file correctly. This could also lead firms to rely on heuristics such as previous months' returns and file the same amount again. If unable to file same amount again, due to a notch, firms could deviate either upwards or downwards (Tourek, 2022). We argue that record keeping and this form of heuristics play a minor role in the observed bunching pattern for two main reasons. First, the share of *targeted past payments* is uncorrelated with the phenomenon of bunching above thresholds (cf. Table 3). Second, our survey shows that 89% firms are able to keep records of at least some form (cf. Table 4).

6.3 Mistakes.

One could also consider the possibility that firms simply make mistakes. This could happen when firms are inattentive to the tax schedule or do not understand it

correctly. Almunia et al. (2023) for example argue that in Uganda, a substantial share of VAT returns are wrongly filed because firms are confused.

We provide five pieces of evidence to show that such an explanation is insufficient to rationalize bunching above thresholds. First, we see the cross sectional bunching pattern for all periods from January 2017 to December 2018 (see Figures A.7 and A.8). One would assume that if firms were simply confused, there would be at least some learning over 24 filing periods, which we do not find. Second, we do not observe firms to directly 'correct' their turnover from above the threshold to below the threshold in the next filing period, which would be the natural reaction if bunching above would have constituted a mistake (see Figure A.9). Third, for the sample of firms which are observed both above a threshold as well as below a threshold at different points in time, we document that 58% of these firms bunch above a threshold after they have already bunched below a threshold in a previous tax period. This is clearly inconsistent with the idea that bunching above a threshold simply constitutes a mistake. Fourth, in our survey, we elicited that most firms are aware of the correct tax rate, speaking against the channel of mere tax illiteracy. Finally, the focus on round number tax liabilities suggest that firms are aware of the implications of filing certain turnover for the tax liability.

7 Conclusion

This paper shows that bargaining over tax payments is an important feature of tax collection in lower income countries. The empirical setting is Zambia but we argue that the factors driving this behaviour are similar in many other low and lower-middle income countries and thus, our results apply more generally.

We study firms subject to turnover taxation by analyzing administrative data on the universe of turnover tax filings in Zambia and establishing two novel empirical facts. First, we find strong and sharp bunching *above* tax schedule discontinuities which is a strictly dominated choice in standard models. Second, we find strong bunching at odd turnover amounts, which imply round number liabilities. These observations are at odds with predictions from standard models of tax compliance, but can be rationalized when interpreting tax payments as bargaining outcomes between taxpayers and tax officials.

We gather several pieces of evidence from data on tax audits, our own survey of more than 500 firms as well as a randomized survey experiment which reject

competing explanations for the observed bunching patterns. Finally, we propose a simple theoretical framework of tax compliance, which explains how and when bargaining occurs and rationalizes both empirical facts. It shows that bargaining over taxes leads to pareto-improvements for both taxpayers and the state as long as state capacity is sufficiently low. As countries develop state capacity, bargaining over taxes becomes obsolete.

Overall, our results inform the debate of how the circumstances of lower income countries shape the way (tax-)administration works. We show how puzzling facts in the tax data can be rationalized when accounting for the presence of bargaining as a mode of tax compliance and –enforcement. While the fear might be that such a system generously invites corruption, we argue that bargaining over taxes benefits both taxpayers and tax authority. Hence, shutting it off has welfare implications, the quantification of which we consider an important avenue for future research. Rather, our study suggests that considering this mechanism yields important insights when formulating policy recommendations. In particular, the official tax schedule is less important than one might think. Reforms of the latter may therefore have little impact on revenue. Instead, other policies which could increase bargaining outcomes may be more promising in this regard. As such, lowering the VAT threshold and thereby exposing more firms to third-party reporting constitutes one example.

References

Albers, Wulf and Gisela Albers, "On the prominence structure of the decimal system," in "Advances in Psychology," Vol. 16, Elsevier, 1983, pp. 271–287.

Allingham, Michael G and Agnar Sandmo, "Income tax evasion: A theoretical analysis," *Taxation: critical perspectives on the world economy*, 1972, *3* (1), 323–338.

Almunia, Miguel, Jonas Hjort, Justine Knebelmann, and Lin Tian, "Strategic or confused firms? evidence from âmissingâ transactions in Uganda," *Review of Economics and Statistics*, 2023, pp. 1–10.

Aman-Rana, Shan and Clement Minaudier, "Spillovers in State Capacity Building: Evidence from the Digitization of Land Records in Pakistan," 2024.

_ , _ , and Sandip Sukhtankar, "Informal fiscal systems in developing countries," Technical Report, National Bureau of Economic Research 2023.

Anagol, Santosh, Allan Davids, Benjamin B Lockwood, and Tarun Ramadorai, *Diffuse Bunching with Frictions: Theory and Estimation*, Centre for Economic Policy Research, 2022.

Angelucci, Manuela and Giacomo De Giorgi, "Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption?," *American economic review*, 2009, *99* (1), 486–508.

ATAF, African Tax Administration Forum, "African Tax Outlook 2018," Technical Report, African Tax Administration Forum 2018.

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

Balan, Pablo, Augustin Bergeron, Gabriel Tourek, and Jonathan L Weigel, "Local elites as state capacity: How city chiefs use local information to increase tax compliance in the democratic republic of the Congo," *American Economic Review*, 2022, *112* (3), 762–797.

Basurto, Maria Pia, Pascaline Dupas, and Jonathan Robinson, "Decentralization and efficiency of subsidy targeting: Evidence from chiefs in rural Malawi," *Journal of Public Economics*, 2020, *185*, 104047.

Besley, Timothy and Torsten Persson, "Taxation and development," in "Handbook of Public Economics," Vol. 5, Elsevier, 2013, pp. 51–110.

_ and _ , "Why do developing countries tax so little?," *Journal of Economic Perspectives*, 2014, *28* (4), 99–120.

Best, Michael Carlos, Anne Brockmeyer, Henrik Jacobsen Kleven, Johannes Spinnewijn, and Mazhar Waseem, "Production versus revenue efficiency with limited tax capacity: theory and evidence from Pakistan," *Journal of Political Economy*, 2015, *123* (6), 1311–1355.

Blattman, Christopher, Julian Jamison, Tricia Koroknay-Palicz, Katherine Rodrigues, and Margaret Sheridan, "Measuring the measurement error: A method to qualitatively validate survey data," *Journal of Development Economics*, 2016, *120*, 99–112.

Carrillo, Paul, Dave Donaldson, Dina Pomeranz, and Monica Singhal, "Ghosting the Tax Authority: Fake Firms and Tax Fraud," Technical Report, National Bureau of Economic Research 2022.

Fellner, Gerlinde, Rupert Sausgruber, and Christian Traxler, "Testing enforcement strategies in the field: Threat, moral appeal and social information," *Journal of the European Economic Association*, 2013, *11* (3), 634–660.

Gadenne, Lucie and Monica Singhal, "Decentralization in developing economies," *Annu. Rev. Econ.*, 2014, *6* (1), 581–604.

Gaspar, Vitor, Laura Jaramillo, and Philippe Wingender, "Tax Capacity and Growth: Is there a Tipping Point?," *IMF Working Paper*, 2016, *234* (16).

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

Hindriks, Jean, Michael Keen, and Abhinay Muthoo, "Corruption, extortion and evasion," *Journal of Public Economics*, 1999, *74* (3), 395–430.

Hoy, Christopher, Thiago Scot, Alex Oguso, Anna Custers, Daniel Zalo, Ruggero Doino, Jonathan Karver, and Orgeira Nicolas Pillai, "Trade-offs in the Design of Simplified Tax Regimes: Evidence from Sub-Saharan Africa," *World Bank Policy Research Paper No. 10909*, 2024.

International Monetary Fund, "Stepping up Domestic Resource Mobilization: a new joint initiative from the IMF and World Bank," *https://www.imf.org/-/media/Files/Research/imf-and-g20/2024/domestic-resource-mobilization.ashx (Last Accessed 24/10/2024)*, 2024.

Janssen, Maarten CW, "Rationalizing focal points," *Theory and Decision*, 2001, *50*, 119–148.

_ , "On the strategic use of focal points in bargaining situations," *Journal of Economic Psychology*, 2006, *27* (5), 622–634.

Khan, Adnan Q, Asim I Khwaja, and Benjamin A Olken, "Tax farming redux: Experimental evidence on performance pay for tax collectors," *The Quarterly Journal of Economics*, 2016, *131* (1), 219–271.

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

_ **and Mazhar Waseem**, "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan," *The Quarterly Journal of Economics*, 2013, *128* (2), 669–723.

_ , **Claus Thustrup Kreiner, and Emmanuel Saez**, "Why can modern governments tax so much? An agency model of firms as fiscal intermediaries," *Economica*, 2016, *83* (330), 219–246.

_ , **Martin B Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez**, "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark," *Econometrica*, 2011, *79* (3), 651–692.

Mascagni, Giulia and Christopher Nell, "Tax compliance in Rwanda: Evidence from a message field experiment," *Economic Development and Cultural Change*, 2022, *70* (2), 587–623.

_ , **Fabrizio Santoro, Denis Mukama, John Karangwa, and Napthal Hakizimana**, "Active ghosts: Nil-filing in Rwanda," *World Development*, 2022, *152*, 105806.

Ministry of Labour and Social Security, "An Analysis of the informal economy in Zambia," Technical Report, Central Statistical Office, Lusaka 2018.

Okunogbe, Oyebola and Gabriel Tourek, "How Can Lower-Income Countries Collect More Taxes? The Role of Technology, Tax Agents, and Politics," *Journal of Economic Perspectives*, 2024, *38* (1), 81–106.

_ **and Victor Pouliquen**, "Technology, taxation, and corruption: evidence from the introduction of electronic tax filing," *American Economic Journal: Economic Policy*, 2022, *14* (1), 341–372.

Olken, Benjamin A and Monica Singhal, "Informal taxation," *American Economic Journal: Applied Economics*, 2011, *3* (4), 1–28.

Pope, Devin G, Jaren C Pope, and Justin R Sydnor, "Focal points and bargaining in housing markets," *Games and Economic Behavior*, 2015, *93*, 89–107.

Saez, Emmanuel, "Do taxpayers bunch at kink points?," *American economic Journal: economic policy*, 2010, *2* (3), 180–212.

Schelling, Thomas, *The Theory of Conflict*, Harvard University Press, 1960.

Shimeles, Abebe, Daniel Zerfu Gurara, and Firew Woldeyes, "Taxman's dilemma: coercion or persuasion? Evidence from a randomized field experiment in Ethiopia," *American Economic Review*, 2017, *107* (5), 420–424.

Slemrod, Joel, "Buenas notches: lines and notches in tax system design," *eJTR*, 2013, *11*, 259.

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

Tourek, Gabriel, "Targeting in tax behavior: Evidence from Rwandan firms," *Journal of Development Economics*, 2022, *158*, 102911.

Udry, Christopher, "Risk and insurance in a rural credit market: An empirical investigation in northern Nigeria," *The Review of Economic Studies*, 1994, *61* (3), 495–526.

United Nations, "The Least Developed Countries Report," Technical Report, United Nations Conference on Trade and Development, geneva 2023.

World Bank, "Risk-based tax audits: Approaches and country experiences," Technical Report, World Bank Publications 2011.

_ , "Small and Medium Enterprises (SMEs) Finance," *https://www.worldbank.org/en/topic/smefinance (Last Accessed 28/07/2024)*, 2019.

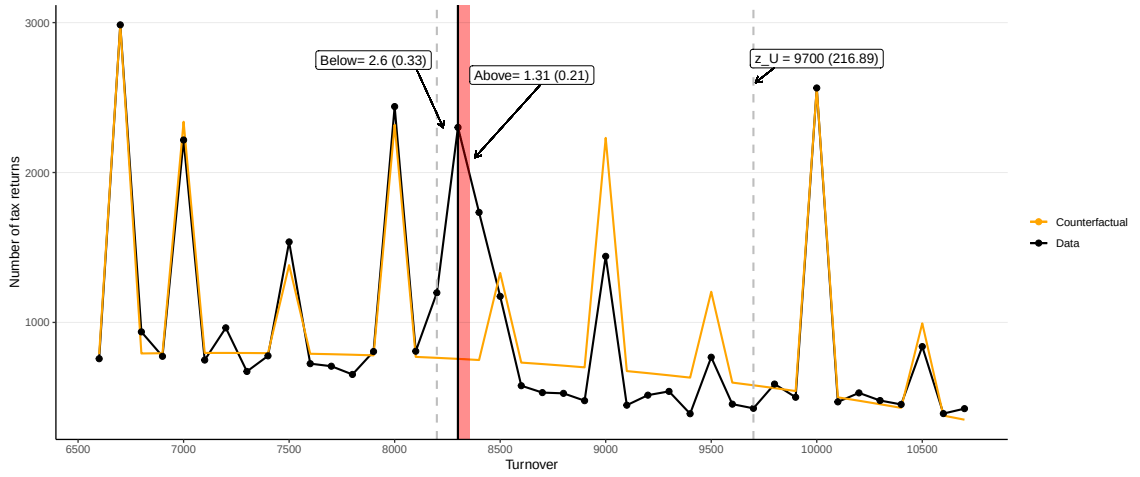
Zambia Revenue Authority, "Tax Statistics in Zambia 2021," Technical Report, Central Statistical Office, Lusaka 2022.

A Anatomy of bunching *above*

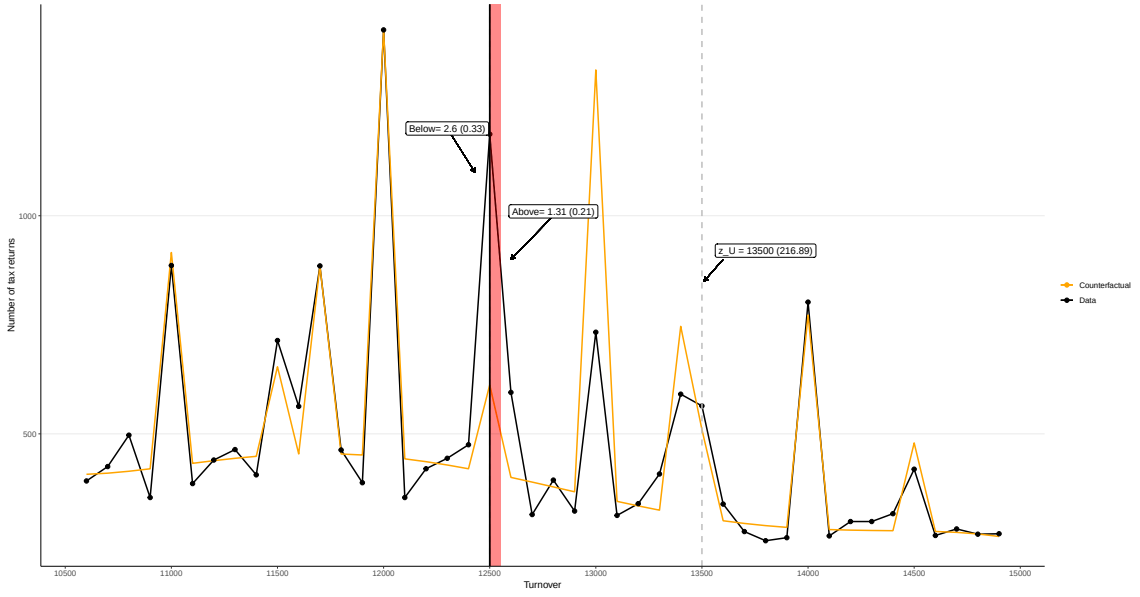
This section complements section 4 by further investigating the phenomenon of bunching *above* the threshold. Figures A.1 and A.2 plot the bunching figures for the other thresholds. Figures A.3 and A.4 plots the distribution of tax returns before and after the notched schedule has been in place, respectively. Figures A.5 and A.6 plot bunching responses with different counterfactual distributions. Figures A.7 and A.8 show that the cross sectional distribution is consistent across months. Figure A.9 shows that bunching firms do not immediately bunch below in the following periods. In Figure A.11, we show that most firms only bunch *above* once. Tables A.1 and A.2 show the heterogeneity of bunching above by sector and tax office, respectively.

Figure A.1: Bunching around thresholds 8300 and 12500

(a) Bunching at 8300



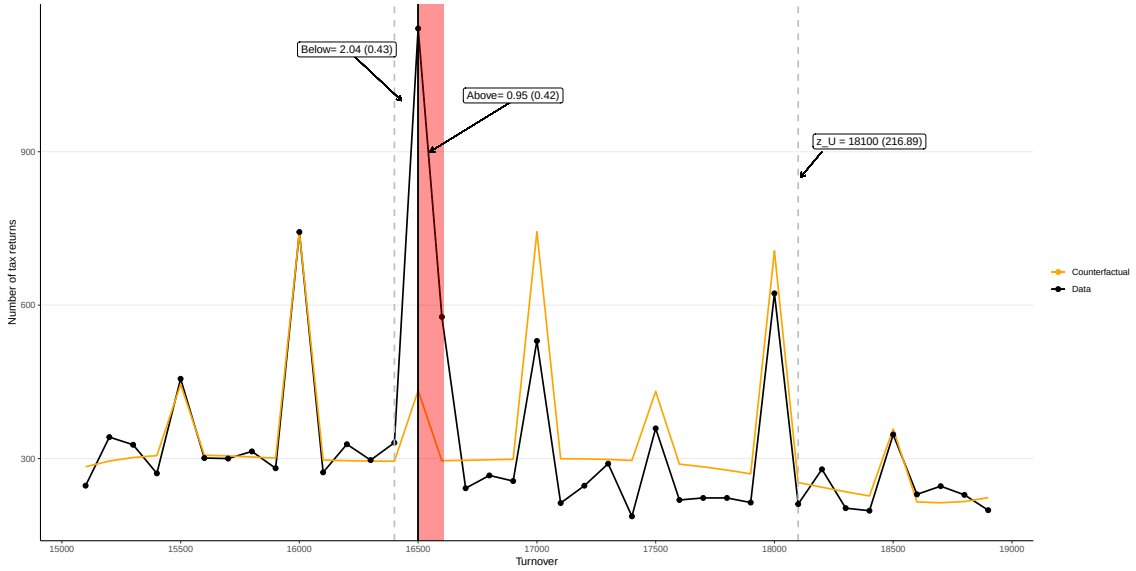
(b) Bunching at 12500



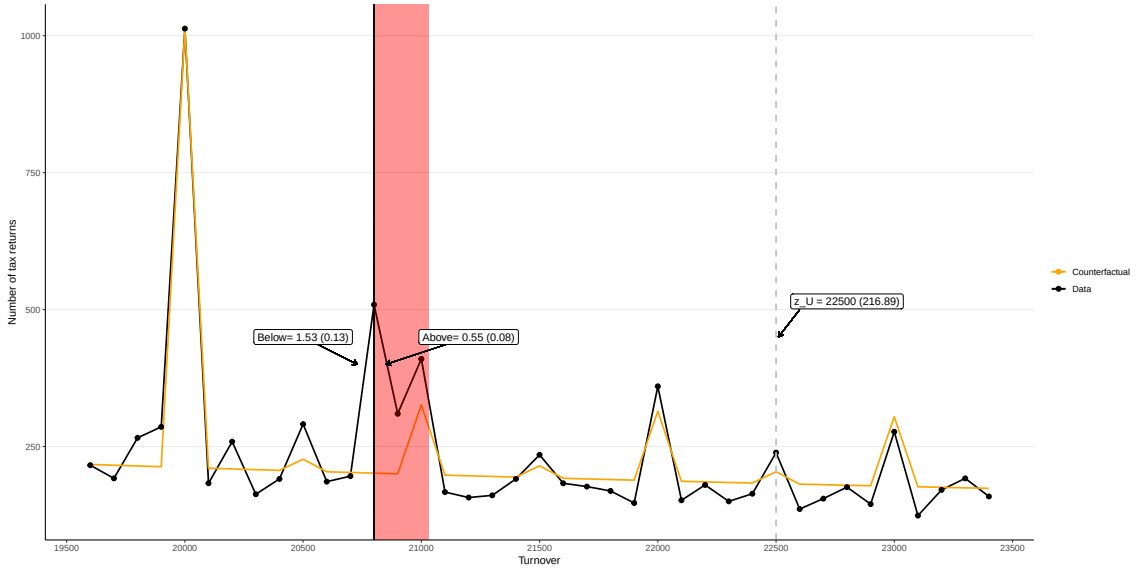
Notes: This figure plots the results of estimating bunching at the 8300 and 12500 thresholds (panel (a) and (b), respectively). The black lines depict the empirical densities. The yellow lines depict the counterfactual densities as estimated by Eq. (1), accounting for round turnover amounts as well as round payment amounts.The black solid vertical lines mark the threshold at which tax liability increases discretely i.e. the notch. The red area depicts the dominated range. The grey dashed vertical lines depict the lower and upper bounds of the omitted region z_L, z_U. Estimates of bunching below and above the threshold are derived from Eq. (2) & (3) and compare the counterfactual to the empirical density. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. Data source: ZRA. Years: 2017, 2018

Figure A.2: Bunching around thresholds 16500 and 20800

(a) Bunching at 16500

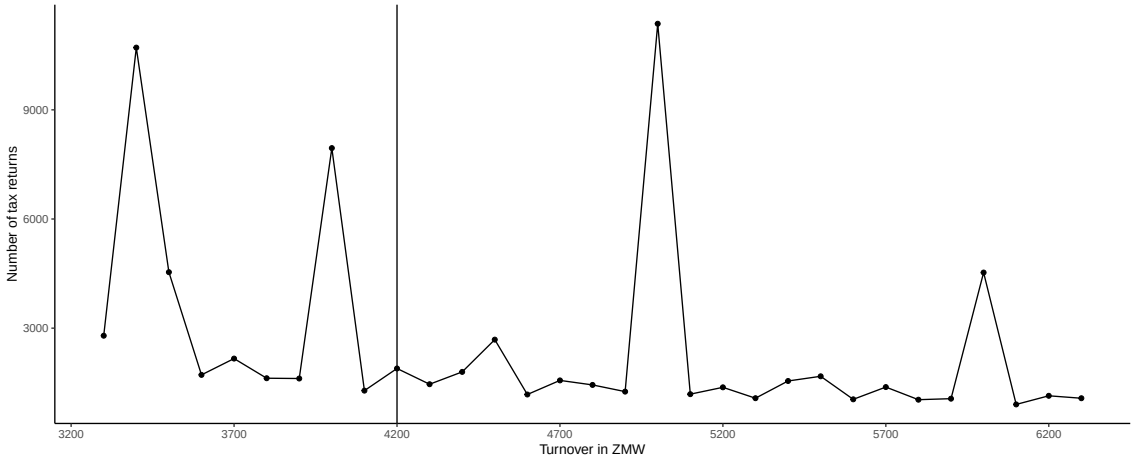


(b) Bunching at 20800



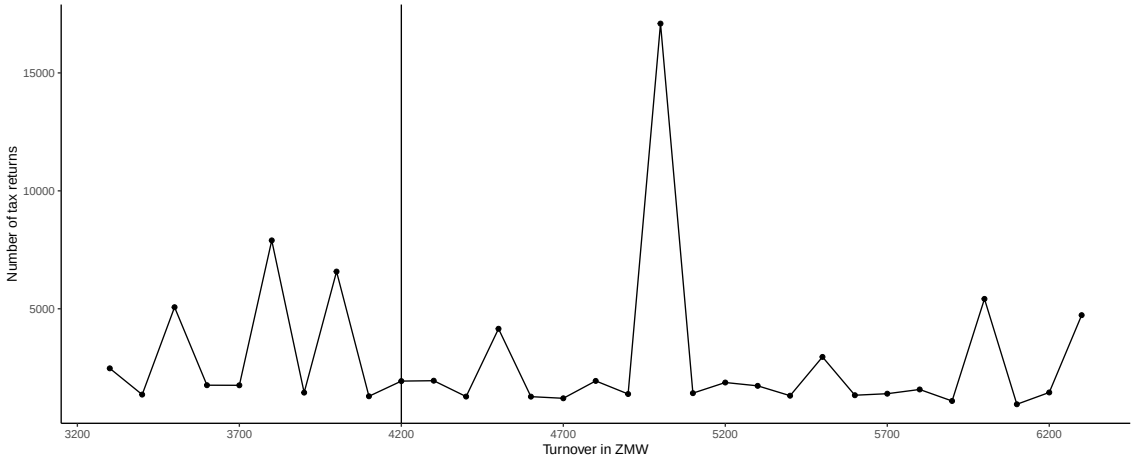
Notes: This figure plots the results of estimating bunching at the 16500 and 20800 thresholds (panel (a) and (b), respectively). The black lines depict the empirical densities. The yellow lines depict the counterfactual densities as estimated by Eq. (1), accounting for round turnover amounts as well as round payment amounts. The black solid vertical lines mark the threshold at which tax liability increases discretely i.e. the notch. The red area depicts the dominated range. The grey dashed vertical lines depict the lower and upper bounds of the omitted region z_L, z_U. Estimates of bunching below and above the threshold are derived from Eq. (2) & (3) and compare the counterfactual to the empirical density. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. Data source: ZRA. Years: 2017, 2018

Figure A.3: Distribution before notched schedule



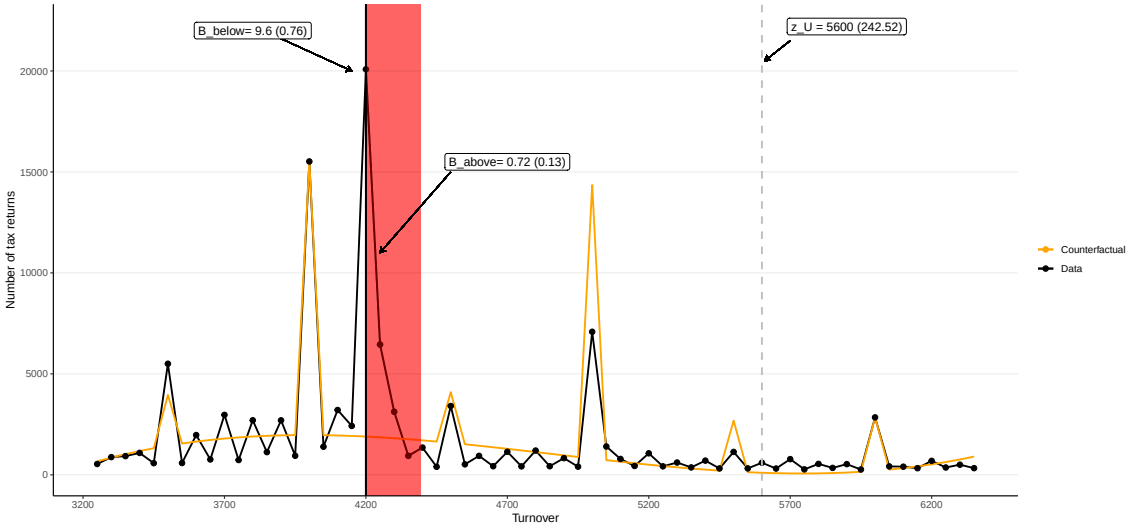
Notes: This figure plots the distribution of tax returns around the threshold before the notched schedule was in place. Data source: ZRA. Years: 2015, 2016

Figure A.4: Distribution after notched schedule



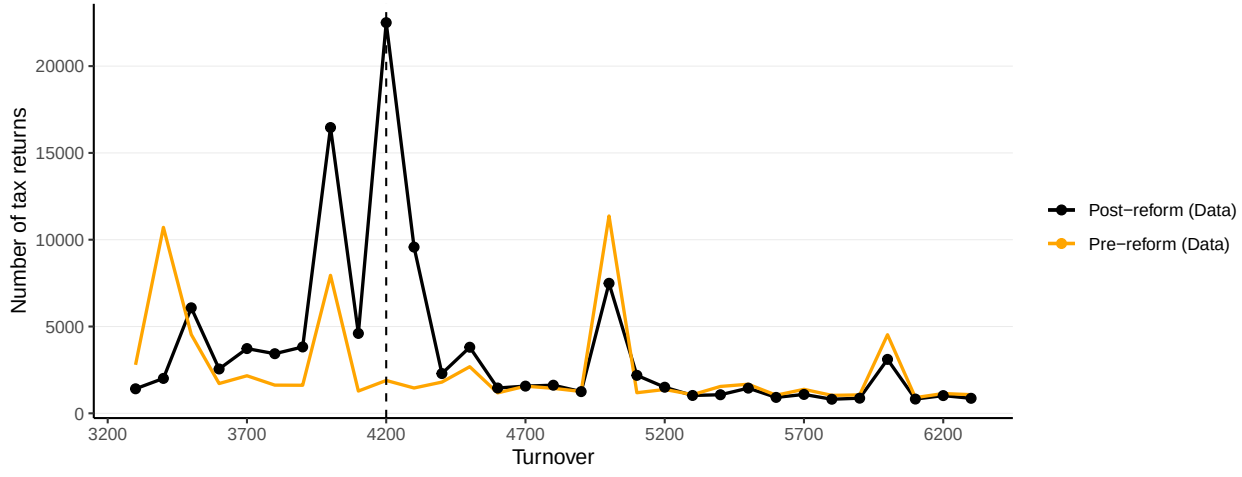
Notes: This figure plots the distribution of tax returns around the threshold after the notched schedule was in place. Data source: ZRA. Years: 2019-2021

Figure A.5: Bunching with binsize of 50



Notes: This figure plots the bunching response using a binsize of 50. Data source: ZRA. Years: 2017-2018

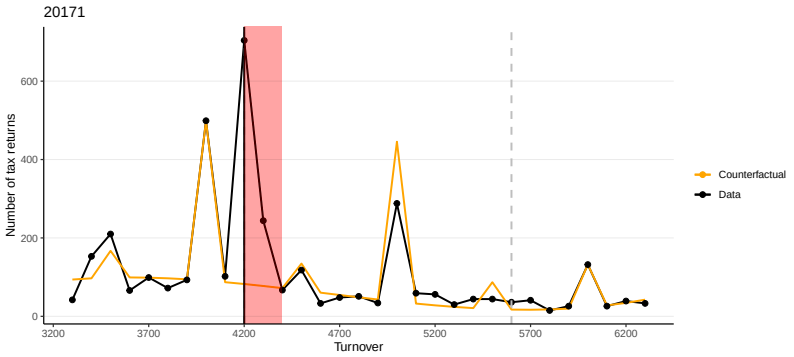
Figure A.6: Bunching with pre-reform data as counterfactual



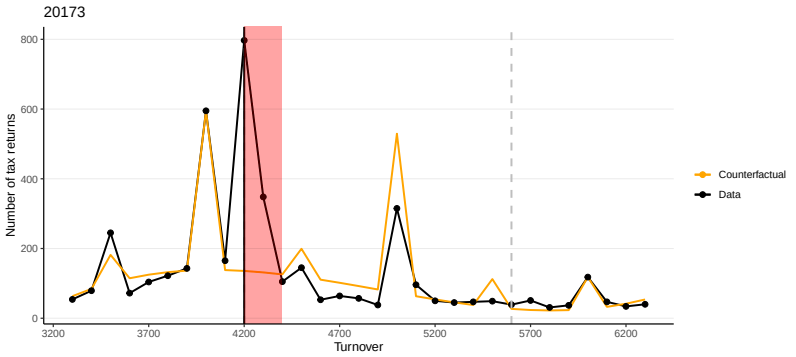
Notes: This figure plots the distribution of tax returns around the threshold before and during the notched schedule was in place. Data source: ZRA. Years: 2015-2018

Figure A.7: Bunching patterns over time 2017

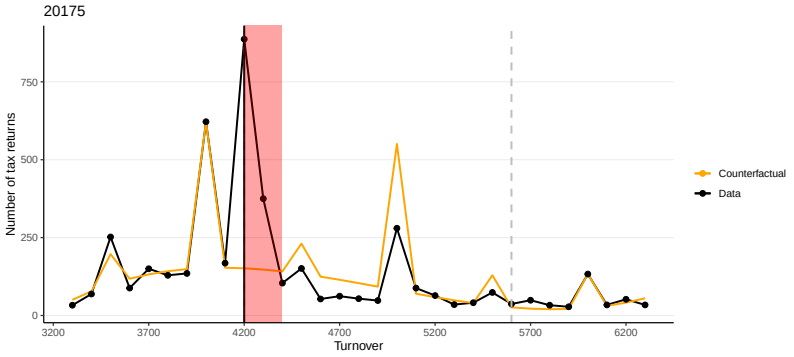
(a) Jan 2017



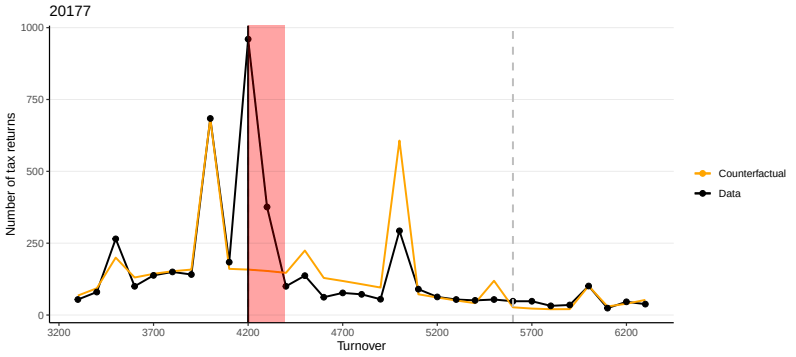
(b) March 2017



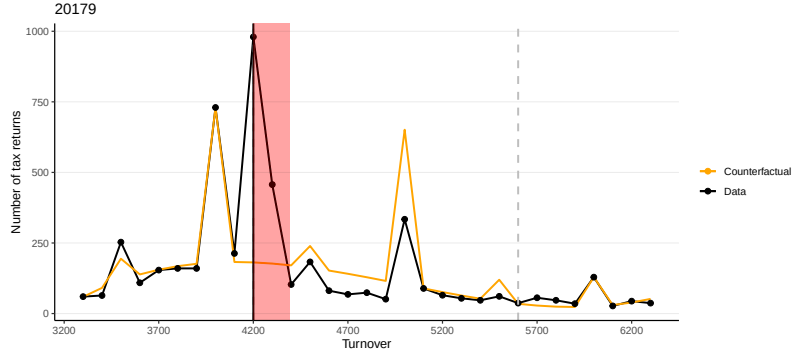
(c) May 2017



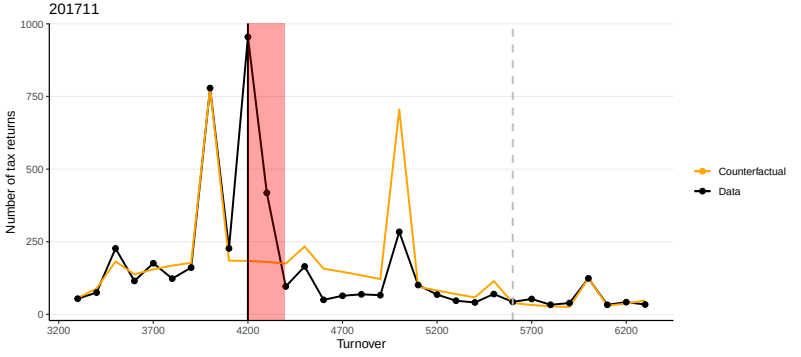
(d) July 2017



(e) Sep 2017

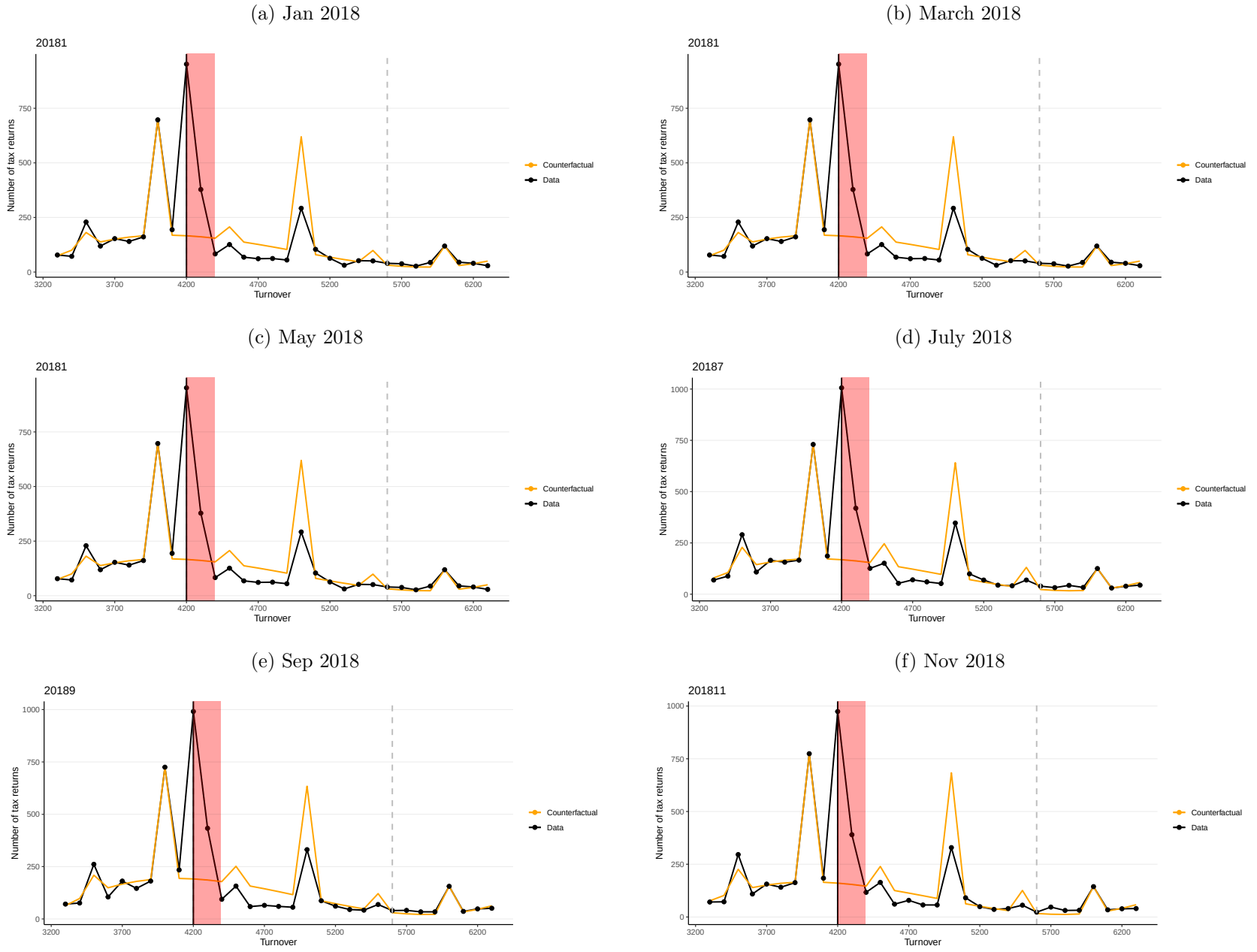


(f) Nov 2017



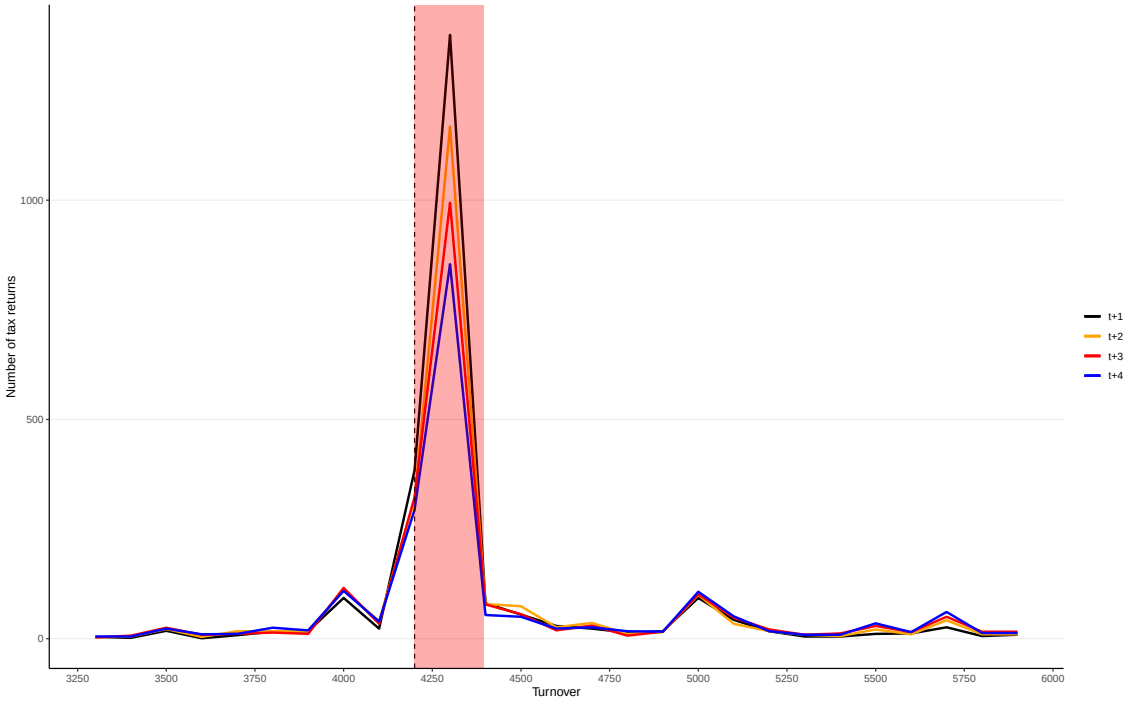
Notes: This figure plots the results of estimating bunching at the 4200 threshold in different months throughout 2017. The black line depicts the empirical density. The yellow line depicts the counterfactual density as estimated by Eq. (1), accounting for round turnover amounts as well as round payment amounts. The black solid vertical line marks the threshold at which tax liability increases discretely i.e. the notch. The red area depicts the dominated range. The grey dashed vertical line depicts the upper bound of the omitted region z_U. Estimates of bunching below and above the threshold are derived from Eq. (2) & (3) and compare the counterfactual to the empirical density. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. Data source: ZRA. Years: 2017

Figure A.8: Bunching patterns over time in 2018



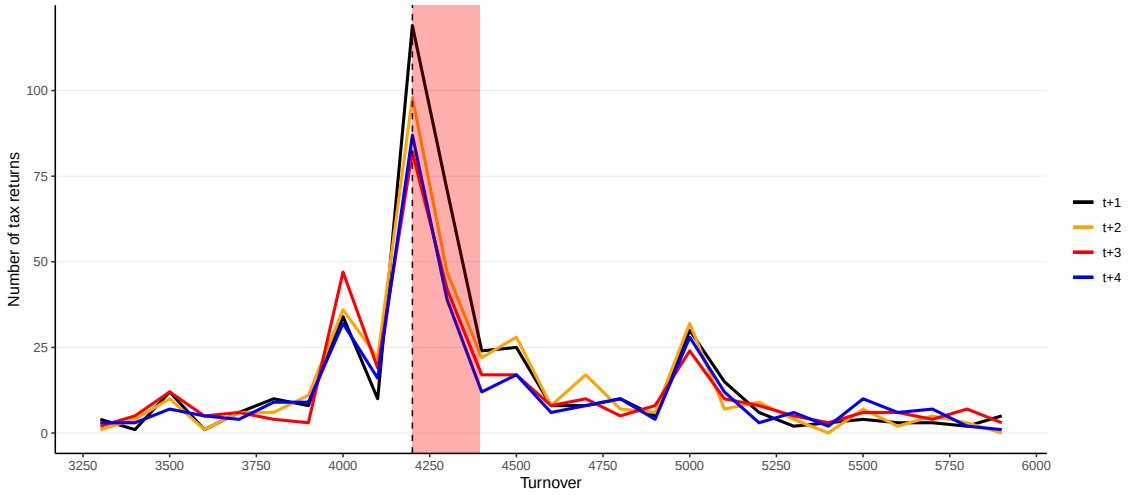
Notes: This figure plots the results of estimating bunching at the 4200 threshold in different months throughout 2018. The black line depicts the empirical density. The yellow line depicts the counterfactual density as estimated by Eq. (1), accounting for round turnover amounts as well as round payment amounts. The black solid vertical line marks the threshold at which tax liability increases discretely i.e. the notch. The red area depicts the dominated range. The grey dashed vertical line depicts the upper bound of the omitted region z_U. Estimates of bunching below and above the threshold are derived from Eq. (2) & (3) and compare the counterfactual to the empirical density. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. Data source: ZRA. Years: 2018

Figure A.9: Distribution of firms after bunching *above*



Notes: This figure plots the distribution of firms that bunched above the threshold at 4200 in a any month in the periods thereafter. The red area depicts the dominated range. Panel (a) shows the distribution one month after, panels (b) and (c) two and three months after having bunched *above*.

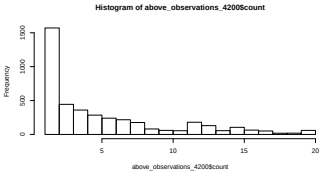
Figure A.10: Distribution of firms after bunching *above* (only once)



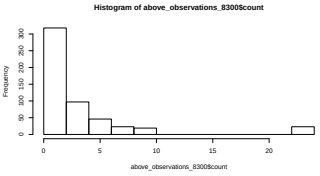
Notes: This figure plots the distribution of firms that bunched above the threshold at 4200 in any month month in the period thereafter. The red area depicts the dominated range.

Figure A.11: Number of times firms bunch *above* the thresholds

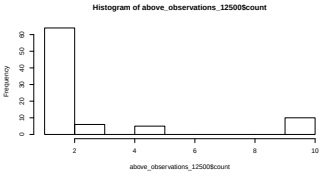
(a) 4200



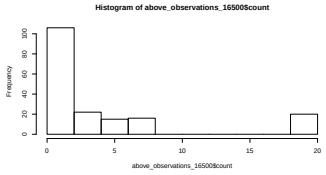
(b) 8300



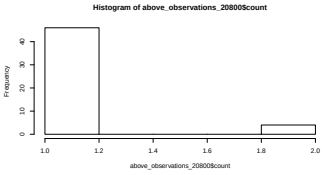
(c) 12500



(d) 16500



(e) 20800



Notes: This figure plots the distribution of firms bunching above the thresholds according to the frequency they do so. For example, panel (a) shows that over 1500 firms which bunch above the 4200 threshold do so only once. Data source: ZRA. Years: 2017, 2018

Table A.1: Bunching above by sector

	Dependent variable:	
	Ever bunched above	
Sector (reference: financial activities)	(1)	(2)
Accommodation and food service activities	**−0.020**	**−0.024**
	(0.035)	(0.035)
Administrative and support service activities	**−0.024**	**−0.027**
	(0.031)	(0.032)
Agriculture, forestry and fishing	**−0.014**	**−0.018**
	(0.031)	(0.032)
Construction	**0.003**	**0.0002**
	(0.032)	(0.032)
Education	**−0.011**	**−0.013**
	(0.037)	(0.038)
Electricity, gas, steam and air conditioning supply	**−0.043**	**−0.031**
	(0.052)	(0.054)
Human health and social work activities	**0.009**	**0.006**
	(0.042)	(0.042)
Information and communication	**−0.043**	**−0.046**
	(0.046)	(0.046)
Manufacturing	**−0.029**	**−0.032**
	(0.033)	(0.033)
Mining and Quarrying	**0.007**	**−0.044**
	(0.042)	(0.084)
Other service activities	**−0.021**	**−0.023**
	(0.029)	(0.029)
Professional, scientific and technical activities	**−0.029**	**−0.032**
	(0.033)	(0.033)
Real estate activities	**−0.023**	**−0.029**
	(0.034)	(0.035)
Transportation and storage	**−0.019**	**−0.023**
	(0.031)	(0.031)
Water supply; sewerage, waste management and remediation	**−0.043**	**−0.046**
	(0.084)	(0.084)
Wholesale and retail trade; repair of motor vehicles and motorcycles	**−0.015**	**−0.018**
	(0.029)	(0.029)
Taxoffice FE		✓
# Firms	22,361	22,361
R^2	0.001	0.002

Notes: This table shows the results from regressing an indicator variable of whether a firm has ever bunched *above* a threshold on sector indicator variables. The reference sector is "Financial and insurance activities". Standard errors are in parantheses. Not shown, but small and insignificant are the estimated coefficients for the sectors: "Activities of extraterritorial organizations and bodies", "Activities of households as employers; undifferentiated goods- and services- producing activities of households for own use", "Arts, entertainment and recreation", "Public administration and defence; compulsory social security" and non-classified sectors. Data source: ZRA. Years: 2017-2018.

Table A.2: Bunching above by taxoffice

	Dependent variable:	
	Ever bunched above	
Tax office (reference: Lusaka)	(1)	(2)
Central Province	**0.0004**	**0.001**
	(0.006)	(0.006)
Chingola	**0.003**	**0.004**
	(0.008)	(0.008)
Choma	**−0.010**	**−0.009**
	(0.007)	(0.007)
Eastern Province	**0.015****	**0.015****
	(0.007)	(0.007)
Kitwe	**−0.008**	**−0.008**
	(0.005)	(0.005)
Livingstone	**0.003**	**0.002**
	(0.009)	(0.009)
Luapula Province	**0.011**	**0.011**
	(0.010)	(0.010)
Muchinga Province	**−0.004**	**−0.004**
	(0.011)	(0.011)
Ndola	**0.006**	**0.007**
	(0.005)	(0.005)
Northern Province	**0.003**	**0.003**
	(0.009)	(0.009)
Northwestern Province	**−0.005**	**−0.005**
	(0.008)	(0.008)
Western Province	**0.003**	**0.004**
	(0.009)	(0.009)
Small Taxpayer Offices (Mining– , Non Mining– , VAT North combined)	**0.036**	**0.057**
	(0.033)	(0.085)
Sector FE		✓
# Firms	22,361	22,361
R^2	0.001	0.002

Notes: This table shows the results from regressing an indicator variable of whether a firm has ever bunched *above* a threshold on tax office indicator variables. The reference tax office is "Lusaka". Standard errors are in parantheses. Data source: ZRA. Years: 2017-2018.

B Bunching at round number tax liabilities.

This Section relates to the analysis in Section 4.1 and provides further evidence on bunching at round number tax liabilities. Table B.1 provides the results from estimating Eq. (4) (i.e. round number tax liabilites which *do not coincide* with round number turnover amounts, as graphically depicted in Figure 4), pooling all tax returns from 2015-2016. Bootstrapped standard errors are in parentheses. Tables B.2 and B.3 show results for the same exercise in only 2015 and 2016, respectively. Further, in Figure B.1 we plot the distribution of deviations from tax liabilities which imply multiples of 10. We do so by creating tax liability bins of size 10, with the multiple of 10 as the mid point, e.g., [5,15]. Then, we calculate how far away each tax return is from its own bin's midpoints. We do so for all bins, for bins with round turnover amounts as the mid point and for those with only round liability amounts as the mid point.

Table B.1: Bunching at round number tax liabilities 2015-2016

	Liabilities					
	40	50	70	80	100	110
B	**9.54**	**30.59**	**11.21**	**18.15**	**10208.6**	**26.79**
	(0.44)	(0.01)	(1.26)	(5.51)	(11.87)	(49.49)
	130	140	160	170	190	200
B	**6.08**	**4.46**	**6.66**	**7.03**	**3.83**	**56.69**
	(0.38)	(0.31)	(0.45)	(0.58)	(0.41)	(5.05)
	220	230	250	260	280	290
B	**5.75**	**3.72**	**28.86**	**3.65**	**5.77**	**2.95**
	(0.3)	(0.21)	(1.3)	(0.24)	(0.46)	(0.33)

Notes: This table shows the estimated bunching coefficients for bunching at round number tax liabilities which *do not coincide* with round turnovers (Eq. (4)) and bunching for turnover interval plotted in Figure 4. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. Data source: ZRA. Years: 2015, 2016.

Table B.2: Bunching at round number tax liabilities in 2015

	Liabilities					
	40	50	70	80	100	110
B	**8.10**	**28.16**	**10.10**	**17.1**	**2806.06**	**24.67**
	(0.61)	(1.79)	(1.07)	(9.74)	(2677.99)	(707.57)
	130	140	160	170	190	200
B	**5.82**	**4.90**	**6.57**	**5.89**	**3.26**	**52.51**
	(0.36)	(0.33)	(0.46)	(0.53)	(0.38)	(5.02)
	220	230	250	260	280	290
B	**5.08**	**3.05**	**26.63**	**4.13**	**5.36**	**2.81**
	(0.28)	(0.21)	(1.63)	(0.33)	(0.53)	(0.35)

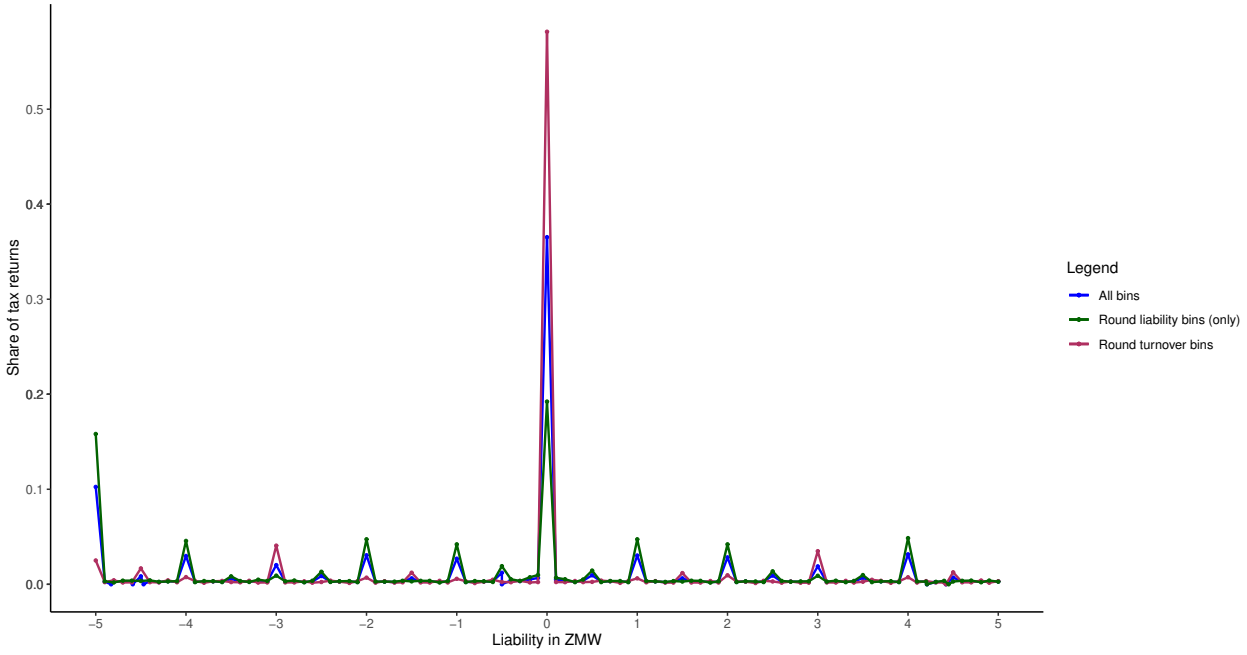
Notes: This table shows the estimated bunching coefficients for bunching at round number tax liabilities which *do not coincide* with round turnovers (Eq. (4)) and bunching for turnover interval plotted in Figure 4. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. Data source: ZRA. Year: 2015

Table B.3: Bunching at round number tax liabilities in 2016

	Liabilities					
	40	50	70	80	100	110
B	**8.95**	**24.13**	**10.14**	**28.9**	**3446.01**	**357.01**
	(0.73)	(1.75)	(1.06)	(8.27)	(224.93)	(558.99)
	130	140	160	170	190	200
B	**7.32**	**4.63**	**7.15**	**8.36**	**4.46**	**63.33**
	(0.42)	(0.29)	(0.46)	(0.64)	(0.43)	(5.26)
	220	230	250	260	280	290
B	**6.31**	**4.62**	**31.44**	**3.08**	**6.02**	**3.61**
	(0.34)	(0.28)	(1.57)	(0.21)	(0.51)	(0.35)

Notes: This table shows the estimated bunching coefficients for bunching at round number tax liabilities which *do not coincide* with round turnovers (Eq. (4)) and bunching for turnover interval plotted in Figure 4. Standard errors are derived from bootstrapping the residuals of the counterfactual density estimation and shown in parentheses. Data source: ZRA. Year: 2016

Figure B.1: Deviations from 10x



Notes: This figure plots the distribution of how much tax returns deviate from the closest turnover amount, which results in a tax liability that is a multiple of 10. *Round liability bins (only)* implies that we only consider tax returns whose closest liability which is divisible by 10 results from odd turnover (e.e. 1333,33). *Round turnover bins* implies that we only consider tax returns also coincides with round turnover as well (e.g. 2000). Data source: ZRA. Years: 2015, 2016.

C Audit Probabilites

This section complements Section 6.1. First, we estimate empirical audit probabilities by turnover. Then, we show theoretically that differential audit probabilities on both sides of the threshold are an unplausible explanation for the phenomenon of bunching *above*.

Empirical audit probabilities around the threshold. We leverage the administrative data on tax audits to test whether there were indeed differential audit probabilities on both sides of the threshold. As most audits in the data refer to more than one tax period, we connect the audits to the tax returns by matching the month the tax return refers to to the audit period end date. This assumes that the last return that has been filed in the tax audit period was the one that triggered the audit. We can therefore infer implicit audit rules, i.e. the empirical probability of being audited conditional on reporting a certain amount of turnover. Figure C.1 plots this empirical audit probability according to turnover bins of size 100 ZMK around the 4200 and 8300 thresholds. As can be seen, conditional audit probabilities were very low for all bins but importantly, there was no substantial difference between the probability of being audited when reporting turnover between $(4100, 4200]$ and $(4200, 4300]$ or $(8200, 8300]$ and $(8300, 8400]$, respectively. Note, that for the other thresholds, there had not been any audits on either side of the threshold throughout 2017 and 2018. This evidence alleviates the concern that bunching *above* thresholds is driven by implicit audit rules. We further address concern from a theoretical perspective in the following.

Tax evasion model and simulation. We consider a firm with turnover z which decides on the tax payment T it plans to make. The firm has quasi-linear utility in z according to

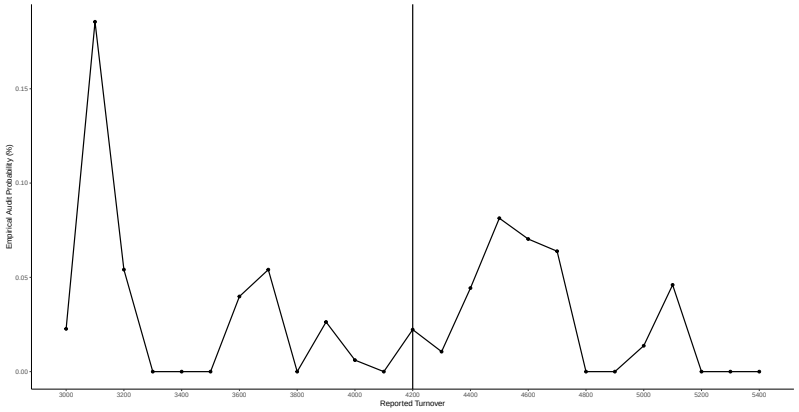
$$U(T, z) = z - T + v(T, z) \tag{15}$$

where $v(T, z)$ is an increasing and concave function in T and governs the decision of how much taxes to pay.

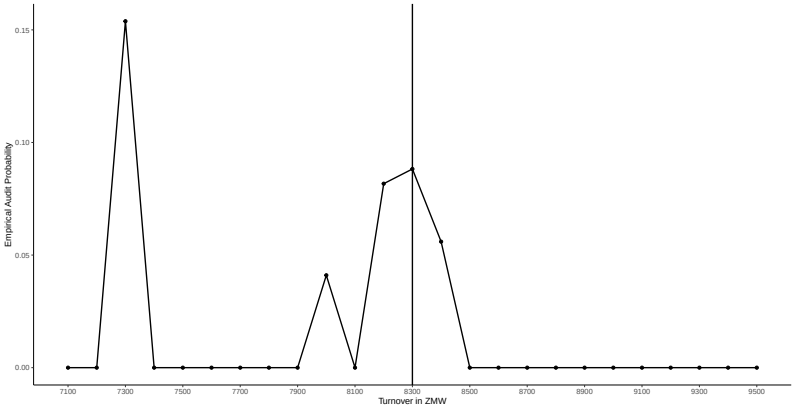
When thinking about how to choose the tax payment T, the firm considers the probability of being caught and the expected penalty. To arrive at a desired tax payment the firm inverts the tax schedule denoted by $T_{schedule}$ and reports turnover accordingly. Thus, for a firm with turnover z that makes a payment T, the probability of being caught is $p(z - T^{-1}_{schedule}(T))$. The firm also considers the penalty which

Figure C.1: Empirical audit probabilities

(a) 4200



(b) 8300

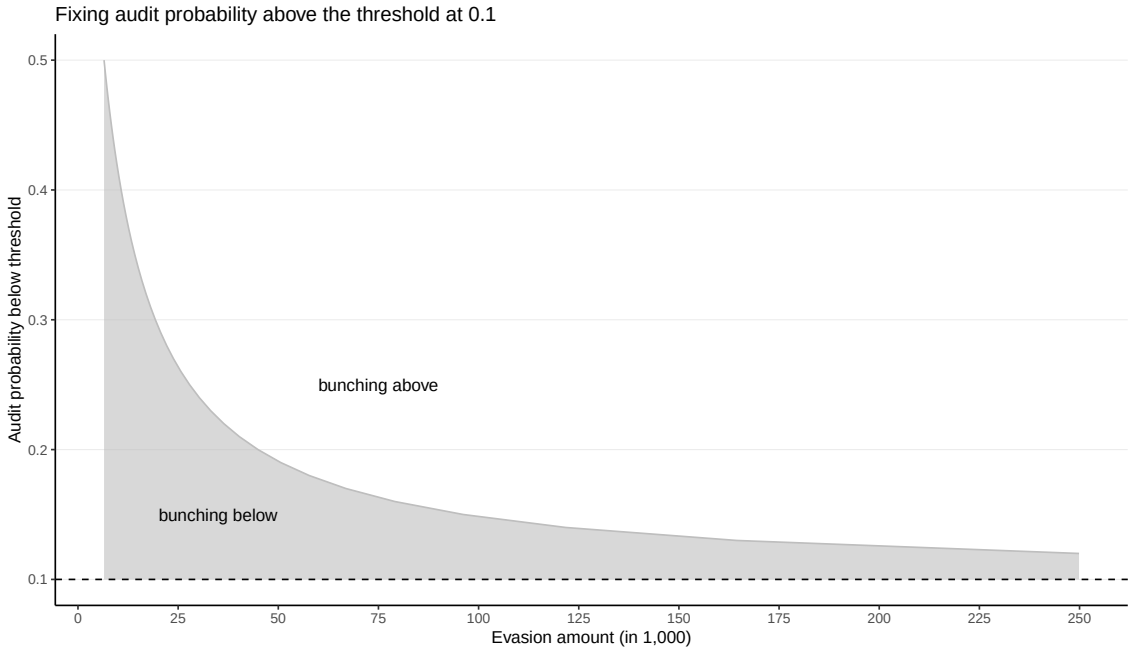


Notes: This figure plots the empirical audit probability (i.e. share of audits per bin) over turnover bins of 100 ZMK. Data source: ZRA. Years: 2017,2018.

we model to be proportional to the evaded amount: $\xi(T(z) - T)$. Implementing the notched schedule from Table 1, we therefore define

$$v(T, z) = p\left(z - \left(\frac{T - F}{t} + \bar{z}\right)\right)(T - T(z))(1 + \xi) \tag{16}$$

Figure C.2: Support for bunching *above* in the audit model



Notes: This figure illustrates under which circumstances firms would choose to bunch *above* – instead of below the threshold due to differential audit probabilities on both sides. The horizontal axis plots the amount of taxes a firm evades while reporting a turnover of 4200 (i.e. bunch below the threshold). When fixing the audit probability above the threshold at 10% for all evaded amounts, the grey area depicts the audit probabilites below the threshold that would support bunching below the threshold in the model sketched above (akin to Kleven et al. (2011)), given true turnover i.e. the amount that is evaded. The white area shows the combinations that support bunching *above*.

such that $U(T, z)$ becomes

$$\begin{aligned}
U(z,T) &= z - T + p\left(z - \left(\frac{T-F}{t} + \bar{z}\right)\right)(T - T(z))(1+\xi) \\
&= z - T + p\left(z - \left(\frac{T-F}{t} + \bar{z}\right)\right)(z - z + T - T(z) + (T - T(z))\xi) \\
&= \left(1 - p\left(z - \left(\frac{T-F}{t} + \bar{z}\right)\right)\right)(z - T) \\
&+ p\left(z - \left(\frac{T-F}{t} + \bar{z}\right)\right)(z - T(z) - (T(z) - T)\xi)
\end{aligned}$$

which is akin to the evasion function from Kleven et al. (2011).

Within this framework, we now investigate which assumptions would generate bunching *above* the threshold. In particular, we consider a firm with true turnover way above the threshold $z >> \bar{z}$ which evades taxes up to the threshold and contemplates choosing to report just below or just above the threshold. If the firm chooses

to report above the threshold instead of below this would imply:

$$U(z, 225) > U(z, 36)$$
$$z - 225 + p\,(z - 4200 + \epsilon)\,(225 - T(z))(1 + \xi) > z - 36 + p\,(z - 4200)\,(36 - T(z))(1 + \xi)$$
$$\frac{189}{1 + \xi}(1 - p(z - 4200)(1 + \theta)) < (p(z - 4200) - p(z - 4200 + \epsilon))(z - \bar{z})t$$

In the Zambian context, one can assume that $\theta \leq 0.095$ (Zambia Revenue Authority, 2022). To investigate under which circumstances, a firm will choose to bunch above, we simulate turnover and audit probabilities. In Figure C.2 the white area shows the combinations of audit probabilities below the threshold (y-axis) and the amount that is evaded (x-axis) that support bunching *above* when the audit probability above the threshold is fixed at 10%.

Overall, these simulation results do not support the hypothesis that bunching *above* thresholds, is the result of differing audit probabilities on both sides of the threshold. For example, even if the audit probability would jump from 10% to 30% at the threshold, only firms evading at least 25,000 (i.e. 83% of turnover) would choose to bunch *above*.

Simulation with risk aversion. We extend the model to allow for different levels of risk aversion. Instead of quasi-linear utility, we now consider a risk-averse firm with CRRA utility of the following form:

$$U(z, T) = \frac{(z - T + v(z, T))^{1-\sigma} - 1}{1 - \sigma}. \tag{17}$$

With this utility function at hand, we calculate under which circumstances the firm's expected utility[19] is larger when bunching above than when bunching below, given that the audit probability is 10 % above the threshold. Let the audit probability above the threshold given z be denoted by $p_{a,z}$. The condition for bunching above is then given by:

$$\mathbf{E}[U(z, 225)] > \mathbf{E}[U(z, 36)]$$
$$p_{a,z} > \frac{(z - 225)^{1-\sigma} - (z - 36)^{1-\sigma} + 0.1((z - 225 - (z - 4200)t(1 + \xi))^{1-\sigma} - (z - 225)^{1-\sigma}}{(z - 36 - (225 + (z - 4200)t - 36)(1 + \xi))^{1-\sigma} - (z - 36)^{1-\sigma}}$$

[19]Note that under the assumption of quasi linear utility, as above, the expectation function is linear, i.e. $\mathbf{E}[z - T + v(z, T)] = z - T + v(z, T)$.

The resulting combinations of $p_{a,z}$ and the evaded amount $z - 4200$ change in comparison to Figure C.2. In particular, the curve becomes flatter. Assuming a risk aversion parameter of $\sigma = 2$, we now find that under the scenario from before, where audit probability triples above the threshold (from 10% to 30%), only firms evading at least 58% of their turnover would choose to bunch above. With higher risk aversion of $\sigma = 3$, this number drops to about 50%.

D Revenue Targets

This section relates to arguments made in Section 5.3. To test the role of tax officials' incentives, we regress an indicator of whether a tax return classifies as *bunching above* on an indicator of whether a tax office has already reached its revenue target or not. The estimating equation reads:

$$\mathbf{1}(\text{bunched above})_{i,t} = \alpha + \beta \mathbf{1}(\text{target reached})_{o,t-1} + \mathbf{X} + \epsilon_i \tag{18}$$

where $\mathbf{X}$ represent fixed effects for tax office (denoted by o), a firm's sector and the threshold at which the tax return bunched (i.e. 4200, 8300,...). Table **??** presents the results.

Table D.1: Tax office revenue targets and *bunching above*

Panel A: All tax returns	*Dependent variable:*			
	Bunched above			
	Whole year		Second half of year	
	(1)	(2)	(3)	(4)
Revenue target reached (0/1)	**−0.0001**	**−0.0001**	**−0.0004*****	**−0.0004*****
	(0.0001)	(0.0001)	(0.0002)	(0.0002)
Observations	1,403,815	1,403,815	751,349	751,349
R^2	0.0001	0.0002	0.0001	0.0002
Taxoffice FE	✓	✓	✓	✓
Sector FE		✓		✓
Baseline mean	**0.0036**	**0.0036**	**0.0039**	**0.0039**
Panel B: Only bunchers	*Dependent variable:*			
	Bunched above			
	Whole year		Second half of year	
	(1)	(2)	(3)	(4)
Revenue target reached (0/1)	**−0.008**	**−0.008**	**−0.012**	**−0.013**
	(0.007)	(0.007)	(0.008)	(0.008)
Observations	25,598	25,598	14,264	14,264
R^2	0.003	0.006	0.004	0.008
Threshold FE	✓	✓	✓	✓
Taxoffice FE	✓	✓	✓	✓
Sector FE		✓		✓
Baseline mean	**0.204**	**0.204**	**0.206**	**0.206**

Notes: This table shows the estimated correlations between the event of a tax return being *above* a threshold in a given month and whether a tax office's yeraly revenue target has been reached in the previous month. The revenue targets are defined as the total turnover tax collections in the previous year. Panel A shows the results when including al tax returns. Panel B shows the results when restricting the sample to tax returns that were bunching either just above (e.g. at 4201) or just below (e.g. at 4200) a threshold. Data source: ZRA. Years: 2016-2018.

E Survey experiment

This section provides more detailed information on the survey experiment, outlined in Section 6.1.

Sampling and randomization. The 517 survey respondents were sampled from market places and business districts across Lusaka. Interviews were held in person by a total of 5 surveyors, who each handled an approximately equal share of the 517 interviews. Firm-owners or people running shops were randomly approached and interviewed. In some cases, the respondent provided suggestions as to where the enumerator could find other firms nearby which could also be interviewed (*snowball approach*). As on each surveying day, new firms were randomly approached, we view the overall sampling as quasi-random. To account for potential endogeneity induced by enumerators relying on such firm networks, we control for enumerator fixed effects in our estimations.

The experimental component was tied to the end of the survey and consisted of randomly providing 3 different pieces of information. Randomization took place at the individual respondent level. That is, in each interview, a random number generator assigned the respondent to 1 of 3 groups, each associated with a different information treatment. This randomization allows us to estimate the causal effect of the information treatment on the answers recorded after the post-treatment. In a later robustness check, we again randomly assigned the respondents into two further groups.

Treatment Messages. The 3 groups mutually exclusively received the following messages:

1. **Control.** "We are now reaching the end of the survey. At this point, we want to share some information with you that the University of Mannheim has gathered. In the year 2022 over 170,000 firms in Zambia were registered under Turnover Tax. Were you aware of the information we just shared with you?"

2. **Audit Treatment.** "We are now reaching the end of the survey. At this point, we want to share some information that the University of Mannheim has gathered on the audits conducted by ZRA. Over the last two years, the total sales which were audited by ZRA increased by almost 20 million Kwacha. This is an increase of 50%. The penalties that had to be paid amounted to 20 million Kwacha. Were you aware of the information we just shared with you?"

3. **Contract Treatment.** "We are now reaching the end of the survey. At this point, we want to share some information with you that the University of Mannheim has gathered. The government of Zambia has comitted to awarding 20% of their business contracts to small firms like yours. This is a large amount of potential business. For securing government contracts, a firm requires clearance from ZRA. Were you aware of this information we just shared with you?"

Outcomes. The aim of the experiment was to test alternative explanations for why firms bunch above a threshold instead of below (besides bargaining). As we, unfortunately, can not link survey respondents to their actual tax declarations, we rely on a stated measure of *bunching above*, as follows. In the beginning of the survey, respondents were asked about the monthly turnover they usually have. Let this stated turnover be denoted by X. Later, but before the treatment occured, the respondent was asked, which tax payment the respondent would find appropriate for a firm with turnover X. Let this preferred tax payment be denoted by Y. After the treatment occured, we confront the respondents with a hypothetical notch, i.e., a region of payments that is not reachable. In this situation, the respondent was asked whether it would rather deviate below or above from its previously stated appropriate payment Y. In particular, the respondent was asked:

> "You have indicated that Y Kwacha would be an appropriate tax payment for a business like yours. Now, please think of a scenario in which the tax schedule does not allow a payment of Y Kwacha. Instead, the tax schedule would only allow payments of either $(1-0.1)\times Y$ Kwacha or $(1+0.1)\times Y$ Kwacha. Which payment would you choose instead?"

If a respondent states to rather pay the larger amount, i.e., $(1+0.1)\times Y$, we interpret this as a propensity to bunch above a threshold. Arguably, this hypothetical situation is similar to the one, a taxpayer would face when filing taxes (without any interaction with tax officials). In the survey, the respondent can simply choose to either pay the larger or the lower amount. In reality, being below or above the threshold is a matter of reporting only a slightly different amount of turnover.

In principle, respondents' answers could be influenced by legal considerations. For example, if firms stated their actual liability as the appropriate payment, i.e., $Y = 0.04 \times X$. Then choosing $(1-0.1)\times Y$ could be viewed as illegal by taxpayers. To check whether these concerns matter for the outcomes, we randomly assign respondents into two groups. We reframe the question and exogenously fix

the respondents' turnover to $X = 10,000$ and their appropriate payment Y to 300 and 500, respectively. In this case, for the first group both payments above and below would be illegal (< 400) while for the second group both answers depict legal amounts (> 400).

Estimating the effects of the randomized information treatments can inform us about the channels which could drive bunching above. The results are presented in Section 6.1 and show that there is no significant effect of either treatment. This supports the notion that bargaining is the most likely explanation for firms bunching above the threshold. Clearly, as in any survey experiment, the outcomes we measured are only stated and do not necessarily coincide wth actions. However, being unable to match survey respondents to their administrative tax records, we consider our approach second-best.

F Proofs

This section provides the proofs for the propositions stated in Section 5.

F.1 Proof of Proposition 1

We start by characterizing the optimal tax payment T^* in the non-cooperative setting, namely the solution to Eq. (6). The first order condition reads:

$$\begin{gathered}(1-p)v'(z-T) = p\xi v'(z-T(z)(1+\xi)+T\xi) \\ \Longleftrightarrow \\ (v')^{-1}(1-p)(z-T) = (v')^{-1}(p\xi)(z-T(z)(1+\xi)+T\xi)\end{gathered}$$

After rearranging, we can write $T^*(p)$ as a linear combination of z and $T(z)$ as follows:

$$T^*(p) = z\underbrace{\left(\frac{(v')^{-1}(1-p)-(v')^{-1}(p\xi)}{(v')^{-1}(1-p)+\xi(v')^{-1}(p\xi)}\right)}_{\equiv \mathbf{K_1}} + T(z)\underbrace{\left(\frac{(v')^{-1}(p\xi)(1+\xi)}{(v')^{-1}(1-p)+\xi(v')^{-1}(p\xi)}\right)}_{\equiv \mathbf{K_2}}. \tag{19}$$

Note that $K_1 < 0$, $K_2 > 1$ and $K_1 + K_2 = 1$.[20] Furthermore, if $\frac{1-p}{p} = \xi$, then $K_1 = 0$ and $K_2 = 1$. As $\frac{1-p}{p}$ is decreasing in p, this means that there is full compliance (i.e. $T^* = T(z)$), if either ξ or p is sufficiently large such that

$$1 = (1+\xi)p. \tag{20}$$

To prove that $T(z)$ lies outside of the bargaining set (T_G, T_F) it is sufficient to show that $T(z) > T_F$. This is equivalent to:

$$\begin{gathered}T(z) > z - v^{-1}\left((1-p)v(z-K_1z-K_2T(z)) + pv(z-(1+\xi)T(z)+\xi K_1 z+\xi K_2 T(Z))\right) \\ \Longleftrightarrow \\ v(z-T(z)) < (1-p)v(z-K_1z-K_2T(z)) + pv(z-(1+\xi)T(z)+\xi K_1 z+\xi K_2 T(Z))\end{gathered}$$

Now, one can see that as p approaches its maximum value $p = \frac{1}{1+\xi}$, the above inequality will become an equality, because $T^*(\frac{1}{1+\epsilon}) = T(z)$, as can be seen from Eq. (19). Thus, to show that the inequality holds strictly for all other audit probabilities, it suffices to show that the right hand side is strictly decreasing in p for all $p < \frac{1}{1+\epsilon}$.

[20] To reconcile the inequalities note our assumption that $\frac{1-p}{p} \geq \xi$ and that $v()$ is concave.

We therefore take the first derivative of the right hand side with respect to p:

$$\begin{aligned}\frac{\partial()}{\partial p} =& -(1-p)T^*(p)v'(z-T^*(p)) - v(z-T^*(p)) \\ &+ \xi p T^{*\prime}(p)v'(z-(1+\xi(T(z)+\xi T^{*\prime}(p)) + v(z-(1+\xi)T(z)+\xi T^*(p)) \\ =& \underbrace{v(z-(1+\xi)T(z)+\xi T^*(p)) - v(z-T^*(p))}_{<0} \\ &+ T^{*\prime}(p)\underbrace{(\xi p v'(z-(1+\xi)T(z)+\xi T^*(p)) - (1-p)v'(z-T^*(p)))}_{=0} \\ &< 0\end{aligned}$$

where the last term is zero by the Envelope Theorem and the optimality condition for $T^*(p)$ from Eq. (6). This concludes the proof of Proposition 1. □

F.2 Proof of Proposition 2.

We start by revisiting the government's optimal choice of p in Eq. (8). The solution reads:

$$c'(p) = T^{*\prime}(p) + (T(z) - T^*(p) - pT^{*\prime}(p))(1+\xi) \tag{21}$$

We multiply both sides of the equation by $\frac{p}{c(p)}$ to express the condition in terms of the cost elasticity κ:

$$\kappa = (T^{*\prime}(p) + (T(z) - T^*(p) - pT^{*\prime}(p))(1+\xi))\,\frac{p}{c(p)}. \tag{22}$$

Taking the total differential of Eq. (22) and rearranging, we get

$$\frac{dp}{d\kappa} = \left[\frac{\Xi}{c(p)^2}\right]^{-1} \tag{23}$$

where Ξ is given by

$$\begin{aligned}\Xi =& (T^{*\prime\prime}(p) - (T^{*\prime}(p) + pT^{*\prime\prime}(p) + T^{*\prime}(p))(1+\xi))pc(p) \\ &+ T^{*\prime}(p) + (T(z) - T^*(p) - pT^{*\prime}(p))(1+\xi))c(p) \\ &- (T^{*\prime}(p) + (T(z) - T^*(p) - pT^{*\prime}(p))(1+\xi))pc'(p)\end{aligned}$$

which, again can be simplified to

$$\begin{aligned}\Xi =&(T^{*\prime}(p) - (T(z) - T^*(p) - pT^{*\prime}(p))(1+\xi))(c(p) - pc'(p)) \\ &+ (T^{*\prime\prime}(p) - (T^{*\prime}(p) + pT^{*\prime\prime}(p) + T^{*\prime}(p))(1+\xi))pc(p) \\ =&(\underbrace{(T^{*\prime}(p)(1-p(1+\xi))}_{\geq 0} + \underbrace{(T(z) - T^*(p))(1+\xi))}_{\geq 0}\underbrace{(c(p) - pc'(p))}_{\leq 0} \\ &+ \underbrace{(T^{*\prime\prime}(p) - (T^{*\prime}(p) + pT^{*\prime\prime}(p) + T^{*\prime}(p))(1+\xi))pc(p)}_{\leq 0} \leq 0\end{aligned}$$

Finally, if $\Xi \leq 0$, it follows that $\frac{dp}{d\kappa} \leq 0$.

We have now established that the optimally chosen audit probability p is increasing as the audit cost elasticity κ decreases. It is left to show that in the limit, as $\kappa \to 1$, the audit probabity is such that there is no room for bargaining i.e. $T^* = T(z) = T^{nc} = T_F = T_G$. To do so, we recall the optimization of the government:

$$\max_p \quad \mathbf{E}[U_G] = T^{nc}(p) - c(p). \tag{24}$$

Clearly, the first part in Eq.(24), T^{nc}, will be maximized when there is full compliance i.e. $T^* = T(z)$. From Eq. (19), we see that this is the case when $p = \frac{1}{1+\epsilon}$. To prove that $p = \frac{1}{1+\epsilon}$ is a maximum also of the whole equation, it is therefore sufficient to show that the second and negative part of Eq. (24) is minimized by $p = \frac{1}{1+\epsilon}$. To do so, we start by considering the optimality condition in the limit i.e. $\kappa \to 1$. Plugging in Eq. (22) and rearranging gives

$$c(p) \underset{\kappa\to 1}{=} (T^{*\prime}(p) + (T(Z) - T^*(p) - pT^{*\prime}(p))(1+\xi))p \tag{25}$$

Plugging in the conjectured optimum $p = \frac{1}{1+\xi}$ yields that

$$\begin{aligned}c(\frac{1}{1+\xi}) &\underset{\kappa\to 1}{=} (T^{*\prime}(\frac{1}{1+\xi}) + (T(Z) - T^*(\frac{1}{1+\xi}) - \frac{1}{1+\xi}T^{*\prime}(\frac{1}{1+\xi}))(1+\xi))\frac{1}{1+\xi} \\ c(\frac{1}{1+\xi}) &\underset{\kappa\to 1}{=} T(Z) - T^*(\frac{1}{1+\xi}) \\ c(\frac{1}{1+\xi}) &\underset{\kappa\to 1}{=} 0.\end{aligned}$$

where the last equality stems from Eq. (19). We have now shown that as $\kappa \to 1$, $p = \frac{1}{1+\xi}$ minimizes the cost function. Hence, $p = \frac{1}{1+\xi}$ will be the optimally chosen audit probability. It follows immediately that $T^* = T(z) = T^{nc} = T_F = T_G$, such that the bargaining set collapses. □