# Marketing Taxation? Experimental Evidence on Enforcement and Bargaining in Malawian Markets

Lucy Martin[*], Brigitte Seim[†], Simon Hoellerbauer[‡], and Luis A. Camacho[§]

**Abstract**

Understanding how to increase state capacity via higher taxation is a core puzzle in state development. Taxation is critical for states to fund key public goods, and taxation may improve state capacity more broadly. This paper argues that tax compliance is fundamentally a community-level, rather than individual-level, phenomenon. Because of this, tax compliance will be easier to achieve, and have more positive downstream effects on the state, when governments target community-level improvements. To test whether it is more effective to focus such interventions on top-down enforcement or bottom-up quasi-voluntary compliance, we ran a multi-arm field experiment in 128 markets in Malawi. We find that the bottom-up intervention, but not the top-down intervention, significantly increased tax compliance. The bottom-up intervention also increased trust in government, tax morale, satisfaction with services, and political engagement. The results show that community-level tax interventions can positively reshape citizen-state relations.

Word count: 11,162

[*] Associate Professor, Department of Political Science, University of North Carolina at Chapel Hill

[†] Associate Professor, Department of Public Policy, University of North Carolina at Chapel Hill

[‡] Postdoctoral Fellow in Data Science and Society, Department of Mathematics and Statistics/Department of Computer Science, Vassar College

[§] Senior Technical Director, Evaluation, Research, and Analytics, Social Impact

# 1 Introduction

How to improve state capacity is one of the core puzzles in political science. State capacity is necessary for governments in developing countries to secure their borders, provide public goods, and develop economically. Yet, decades of failed development efforts suggest that we still know little about how best to improve state capacity in a sustainable way. One dimension of state capacity that has been extremely stubborn to change is taxation. Even while GDP has increased across countries in sub-Saharan Africa and elsewhere, the percent of GDP taken in taxes has remained stable. Without more revenue, governments are unlikely to be able to expand the role of the state, or to escape the current wave of debt defaults (Tilly, 1992; Stasavage, 2011; North and Weingast, 1989; De la Cuesta et al., 2022; Weigel, 2020).

Understanding taxation is especially critical as it can have benefits outside the realm of revenue but still highly relevant to building state capacity. Taxation has the potential to improve perceptions of the state and democracy, lower corruption, and increase citizen political engagement, which could expand state capacity beyond any revenue effects (Ross, 2004; Timmons, 2005; Baskaran and Bigsten, 2013; Brollo et al., 2013; Prichard, 2015; Paler, 2013; Weigel, 2020; De la Cuesta et al., 2022). Yet, despite recent advances, there are still significant gaps in our understanding of taxation and state development in modern low-income states. We argue that one of the primary issues with existing theory and evidence is that it focuses largely on interventions targeting *individuals*, either taxpayers or tax collectors. In contrast, most of the original theories of taxation are actually about *community-level* processes that suggest the need for much broader, community-level interventions, especially in states where existing levels of taxation are low. There has also been insufficient attention to the causal mechanisms underpinning the relationship between taxation and state capacity more broadly, through mediating variables such as bureaucratic effort, citizen political engagement, perceptions of the state, and trust.

Theories of taxation suggest that citizens pay taxes for one of two reasons. First, citizens may pay because the cost of tax compliance is lower than the expected cost of evasion (Allingham and Sandmo, 1972). Second, citizens may pay "quasi-voluntarily", because they feel they are getting something in return, such as public goods or representation (Levi, 1989; North and Weingast, 1989; Bates and Lien, 1985; Prichard, 2015). These theories have, in turn, led to two approaches to increasing tax compliance through individual-level interventions. In one approach, interventions target individual-level beliefs about the costs of evasion through the use of letters, tax collector visits, or other such interventions. In the other approach, researchers have tried to shift citizen beliefs about how taxes are spent, the degree to which others are paying taxes, or the degree to which paying taxes is a duty.

While there have been some successes with increases in enforcement, attempts to increase quasi-voluntary tax compliance have, on average, had little impact. This is especially troubling, as quasi-voluntary compliance is proposed to have significant downstream benefits for citizen-government relations. Additionally, most experimental evidence to date comes from OECD countries, and even experiments run in lower-income countries focus on relatively weak interventions that aim to improve one mechanism or the other, but not both. It is also not clear how well existing theories of tax compliance will travel to a developing-country setting. For example, tax bargaining may only succeed if the state has enough capacity that citizens trust it to keep a bargain. Coercive approaches may likewise require beliefs that the state is sufficiently strong to enforce penalties for non-compliance (for taxpayers) or shirking (for tax collectors). Finally, there is no evidence regarding whether enforcement and bargaining are complementary tactics, or whether using both is actually less effective than one approach alone. The two approaches may also differ in the extent to which they improve perceptions of government legitimacy and trust, key elements to obtain further gains in state capacity.

This paper uses a field experiment, conducted on 128 markets in Malawi, to test the effectiveness of top-down enforcement and bottom-up bargaining approaches to increasing state capacity—in particular tax compliance and its causal mechanisms—in a setting where state capacity is low. In Malawi, as in many sub-Saharan African countries, fees from open-air markets form one of the largest sources of "own-source" revenue for local governments. However, low tax compliance levels limit governments' ability to provide services, and potential taxpayers, in turn, are reluctant to pay taxes until services improve.

From a theoretical standpoint, markets are an optimal location for both tax compliance approaches. As market vendors are gathered in dense, observable locations, local governments should be able to efficiently monitor and enforce revenue collection. Market vendors also meet the preconditions for tax bargaining: they have high collective action potential, are in broad agreement on how tax revenue should be spent, and indicate high theoretical willingness to pay taxes, provided they see benefits in return. The dense nature of markets also makes public goods provision with tax revenues both relatively inexpensive and easily observed by vendors. This suggests that the community-level approaches discussed above should have a high chance of working.

Our experiment consists of two cross-cutting, market-level treatment bundles. The "bottom-up" intervention bundle was designed to improve quasi-voluntary compliance at the community level. It facilitated communication between market vendors and government; constructed new public goods in markets; and increased transparency regarding revenue levels and spending. The construction element makes this study one of the first to change *actual* levels of public services, rather than *perceptions* of or *information* about these services. The "top-down" enforcement bundle improved local governments' ability to collect, track, and manage market revenue collection, focusing again on the market level, rather than individual tax collectors or taxpayers. It improved revenue tracking technology (mobile money); improved government information about taxpayers; and improved tax collectors' incentives to meet revenue targets.

We find that the bottom-up treatment significantly increased tax compliance among vendors. It also led to significant increases in vendors' trust in local government, satisfaction with services, and their belief that paying tax is a duty. This suggests a causal mechanism whereby the bottom-up treatment increased compliance through taxpayer perceptions of state capacity, specifically the criticality of revenue generation for state effectiveness. We also find downstream effects of the bottom-up treatment on political engagement. Taxpayers were more likely to sign a petition for district government requesting more funding for market services and more likely to send text messages to district government demanding a reduction in the over-reach of government power in collecting revenue in this district.

The top-down treatment had a less robust effect on compliance, and no effect on citizen trust and satisfaction. However, it did lead to higher tax collector effort and citizen perceptions of stronger tax enforcement. Critically, treatment effects were restricted to markets that received only one treatment arm. We find much smaller effects in markets that received both treatment arms. We posit that increased enforcement due to the top-down treatment "crowded out" an increase in quasi-voluntary compliance from the bottom-up treatment, leading to a null effect on average.

This paper makes several contributions to the literature on state capacity and taxation. First, our results show that it is possible to increase state capacity via an intervention, and specifically that it is possible to increase tax compliance by jump-starting tax bargaining and a more positive taxation equilibrium. We also show that the quasi-voluntary compliance approach has positive effects on citizen-state relations more generally, while the top-down approach has smaller effects on broader state capacity. This alleviates worries that increasing taxation will simply lead to more coercive governments, and also points to the potential for taxation to improve citizen-state relations and further bolster state capacity.

Second, this experiment is one of the first that treats tax compliance as a community, rather than individual, process. This is critical because many theories of tax compliance rely on community-level variables like the level of public goods provision or beliefs about whether others are also paying. Thus, our research design allows us to test a key element of tax compliance theory that cannot be addressed with experiments that rely on treating individuals, and to examine community-level outcomes such as capacity.

Our intervention is also much stronger than many previous tax experiments: our bundled interventions are designed to fix multiple broken linkages at once. This allows us to provide evidence on the potential for government interventions to increase tax compliance in low-capacity states. It suggests, however, that weak or single-pronged approaches are unlikely to work, as even our extremely strong interventions had limited success.

## 2 Theory

There are two dominant theories of why citizens pay taxes. This section discusses the theoretical framework for each approach, existing evidence, and how we expect each to work when interventions are targeted at the individual or community level. Each of our community-level experimental treatments—described in full in the following sections—was designed to test how a particular theoretical approach to taxation affects both compliance and state capacity more broadly.

*Increasing compliance via top-down enforcement.* Enforcement-based theories of taxation assume taxpayers are strictly economically rational, and will comply when the costs of the tax are lower than the expected costs of noncompliance; these include the probability of detection and the penalty once caught (see, e.g., seminal work by Allingham and Sandmo (1972)). This implies that an intervention that increases the expected costs of tax evasion should increase compliance. Two of the key ways to change the costs of noncompliance are to increase government information about taxpayers (thus decreasing the costs of monitoring) and to improve the incentives of tax collectors to work hard and enforce taxation. These "top-down enforcement" approaches to tax compliance have the potential to increase compliance and revenue, but also to increase bureaucratic capacity and effort more broadly, which in turn will lead to further gains in state capacity.

Field experiments on taxation consistently find support for this approach: increasing the actual or perceived probability of detection and punishment does in fact improve tax compliance (Coleman, 1996; Slemrod, Blumenthal and Christian, 2001; Kleven et al., 2011; Dwenger et al., 2016; Fellner, Sausgruber and Traxler, 2013; Castro and Scartascini, 2015). Critically, two recent studies find similar impacts in Rwanda (Mascagni, Nell and Monkam, 2017) and Ethiopia (Mascagni, Mengistu and Boldeyes, 2018), suggesting that the empirical patterns hold outside of the OECD. All of these experiments target individual taxpayers, typically through sending letters to individual taxpayers: none target more aggregated groups of taxpayers.

Khan, Khwaja and Olken (2016) shows that incentivizing tax collectors can also improve tax compliance through higher tax collector effort; that paper randomized groups of tax collectors to treatment, rather than communities. To the best of our knowledge the primary prior attempt to test community-level changes in enforcement is Weigel and Kabue Ngindu (2023), which finds that visits from tax collectors for a new tax, randomized at the neighborhood level, are effective. However, that paper does not test alternative approaches to collection, and tests implementation of a new tax, rather than attempts to improve compliance with an existing tax.

*Increasing compliance via quasi-voluntary approaches.* The alternate approach to taxation is based on the idea of quasi-voluntary tax compliance. In many settings, citizens appear to pay taxes despite the low probability of punishment (Alm, Jackson and McKee, 1992; Andreoni, Erard and Feinstein, 1998). This can occur if citizens have high "tax morale" and believe that it is their duty (Torgler, 2007). It can also occur under a conditional compliance strategy, in which citizens comply provided they see their funds used on their preferred policies. This "fiscal exchange" can include formal tax bargains that include policy or institutional concessions (Bates and Lien, 1985; Levi, 1989; North and Weingast, 1989), or where there is a clear link between tax payments and public services (Fjeldstad and Therkildsen, 2008).

The evidence on bottom-up approaches is weaker than that on enforcement approaches. Observational studies show that tax compliance is increasing in tax morale, trust in government, satisfaction with public goods provision, and low levels of corruption (Alm, Martinez-Vazque and Torgler, 2006; Levi, Sacks and Tyler, 2009; Picur and Riahi-Belkaoui, 2006). However, experimental efforts to improve voluntary compliance have failed, including treatments that stress citizens' civic duty to pay taxes and provide information about how revenues are spent (Mascagni, Mengistu and Boldeyes, 2018; Castro and Scartascini, 2015; Coleman, 1996; McGraw and Scholz, 1991; Hallsworth et al., 2017). Interventions that aim to increase citizens' perceptions that others are paying taxes have only succeeded where existing levels of tax compliance are relatively high (see, e.g., Coleman (2007)), and have null or negative effects when baseline compliance levels are low (Castro and Scartascini, 2015; Del Carpio, 2013). Critically, most of these experiments again rely on letters to taxpayers or other individual interventions; they do not actually intervene to change community-level beliefs or government behavior.

Citizens may therefore pay taxes because they are compelled to, or because they feel that paying taxes is in line with their values and interests. This suggests that both interventions that increase enforcement as well as interventions that facilitate tax bargaining should increase tax compliance. Where the two approaches differ is in their causal mechanism and downstream effects for other areas of state development and citizen-state relations.

Top-down enforcement approaches to taxation require the state to invest in more taxpayer monitoring, greater bureaucratic capacity, and improved information collection on taxpayers. We expect such an approach to affect compliance and revenue through several causal mechanisms. In particular, we expect increased effort by tax collectors and decreased corruption among them. Corruption in this context could include stealing revenue, but also taking bribes to allow tax evasion. We, note, however, to the extent that such community-level top-down interventions often focus on tax collectors, rather than payers, it will have a limited impact on citizen-state relations. Indeed, if higher enforcement becomes coercion, top-down approaches could *damage* how citizens view the state.

In contrast, bottom-up approaches have the potential to more drastically reshape citizen-state relations, particularly when communities are targeted rather than individuals. If increased tax compliance is the result of higher public goods provision or better bargaining and communication, it has the potential to improve taxpayer satisfaction with government and public services. This in turn can increase government trust, which is critical for the state's ability to promote development and capacity more generally.

A further consequence of the reshaping of citizen-state relations by bottom-up approaches is a change in political engagement. Bottom-up approaches, if they target communities, may improve taxpayers' bargaining position vis-a-vis the government, or may give more agency to taxpayers overall, potentially empowering them to address grievances or appeal to authority both with respect to the taxpaying itself and in the broader political landscape. In addition, if bottom-up approaches result in increased government trust, they may in turn draw taxpayers closer to the government and make them more likely to engage with the state directly on other topics. In other words, it is possible strengthening taxpayers' connection to the state can have downstream effects for other political outcomes.

# 3 Research Context

Malawi is a paradigmatic example of a low-capacity state. Development is low, with an estimated 66.7% of the population multi-dimensionally poor (UNDP, N.d.). Due to low state capacity, Malawi is among the most aid-dependent countries in the world, with aid representing over 37% of the government's budget. Malawians, therefore, perceive government and donor development efforts to be intertwined (Seim, Jablonski and Ahlbäck, 2020).

Local government capacity is especially limited, with significant *de jure* authority over development but *de facto* reliance on central government funding. Own-source revenues are therefore critical for building the capacity of local government, but invariably make up a small percentage of districts' budgets. The largest source of local revenue is market fees. Malawian markets are open-air collections of stalls, with vendors providing a wide range of goods and services. Vendors are charged a fixed fee each day (typically MWK100 - 200, US$0.14-US$0.27), and, in return, the local government is mandated to provide basic market services. Tax collectors visit the market daily to collect fees and give out receipts.[1] Tax collectors then give revenues to the Market Master, who deposits the cash or brings it to the district headquarters.

At baseline, only 27% of market vendors were able to produce a recent tax receipt. In a series of pre-treatment interviews and focus group discussions, vendors and government officials reported two barriers to higher tax compliance. First, vendors are unwilling to pay voluntarily, because they are dissatisfied with market services, believe that tax revenues are co-opted by government officials, and feel excluded from tax collection institutions and processes.[2] Second, low district capacity hampers tax collection in several ways. Information regarding the tax base was limited—some districts lacked even a list of taxed markets in the district, and most had no data on market size. This is compounded by a fee collection process vulnerable to corruption and poorly paid tax collectors: at baseline, tax collectors were paid $0.80 to $1.35 a day, low even in local terms. Significantly, vendors consistently acknowledged that the fee amount is *not* a barrier to compliance.

[1] Staffing needs vary by market size from one part-time collector to 20 full-time collectors.

[2] In one randomly assigned treatment arm, a government-market engagement meeting provided vendors the opportunity to voice their reasons for not paying the daily tax. These three issue categories came up in 80%, 13%, and 16% of the meetings, respectively.

# 4 Research Design

Our field experiment was conducted as one component of a larger development program (see Appendix A for a description). All hypotheses, measures, and analysis were pre-registered with OSF.[3] All stages of the project received IRB approval; see Appendix P for our ethics statement. We first created a list of 209 eligible markets across the program's eight target districts: Balaka, Blantyre, Kasungu, Lilongwe, Machinga, Mulanje, M'mbelwa, and Zomba. We then selected a sample of 128 markets from this list, prioritizing markets with at least 100 vendors. To facilitate block randomization, the number of sampled markets in each district was divisible by four.

Our field experiment randomized two cross-cutting treatment arms at the community (here, market) level. The "bottom-up" (BU) arm was designed to increase vendors' willingness to pay taxes, while the "top-down" (TD) arm was designed to improve government capacity to collect taxes. Each treatment arm had four components, outlined below. Treatments were randomly assigned, stratifying on district and baseline tax compliance levels. This ensured balance along our main outcome, tax compliance. Table 1 shows the resulting four groups of markets.

Table 1: Experimental Design

| | | Treatment 2:<br>Top-Down | |
|---|---|---|---|
| | | Yes | No |
| Treatment 1:<br>Bottom-Up | Yes | Group 1<br>32 markets | Group 2<br>32 Markets |
| | No | Group 3<br>32 Markets | Group 4<br>32 Markets |

[3] For anonymised PAP, see attachment after appendix.

## 4.1 Experimental Treatments

Because this is one of the first experiments to test the top-down and bottom-up mechanisms for improving tax compliance, we bundled several components together for each treatment arm. This approach is in line with pre-experiment fieldwork that showed low compliance was due to multiple concurrent issues. The components of each treatment were rolled out over a one-year period as as part of a larger, five-year USAID program, "LGAP," in the eight sample districts. All other concurrent LGAP components were designed to avoid confounding our analysis. More detailed descriptions of the treatment bundles are presented below, and expanded on in Appendix A. Appendix B reports additional implementation details.

### 4.1.1 Bottom-Up Treatment Bundle

The first intervention bundle was designed to increase vendors' willingness to pay market taxes voluntarily by addressing their concerns over low government transparency, accountability, and service provision. Markets assigned to receive the bottom-up treatments received the following four components:

**Step 1: Facilitate Market Committee Elections and Training.** To facilitate communication between markets and district government, and improve the collective action potential of vendors, market vendor committee elections were held in all markets without a valid market committee (54 markets total). All committees received training on the proper organizational structure for the committee and their roles and responsibilities.

**Step 2: Facilitate Meetings Between Vendors, Market Committees, and Local Government.** Next, districts held public meetings in each market to address vendors' sense of exclusion from the tax system. In addition to vendors and market committees, meetings included political and bureaucratic district representatives, market staff, and group village headmen. The meetings discussed the connection between taxes and market development;

perceived problems with the current market tax system; and market services and priorities. District officials also introduced the final two components of the treatment: the infrastructure projects (Step 3); and the SMS system (Step 4). Vendors then chose their preferred market infrastructure project. Forty-six markets chose a borehole – the others chose a mix of market sheds, water access, electricity, pathways, concrete slabs, and refuse bins.

**Step 3: Jump Start Service Delivery in Markets.** To escape the low services / low compliance equilibrium, all treatment markets received funding for an infrastructure project, selected based on the priorities generated in Step 2. Projects cost approximately US$5,000. Each project was bookended by opening and handover ceremonies, attended by government officials, market committees, and other vendors.

**Step 4: Increase Transparency in Taxation via a SMS System.** To facilitate ongoing communication and transparency between district governments and market vendors, a two-way vendor-government SMS system was introduced during the Step 2 meetings. Seventy-three percent of meeting attendees signed up for the system. Each month, vendors who opted in received a message with information on the previous month's market revenue and how the money was allocated. Vendors could also use the SMS system to report complaints and grievances about local government service delivery; these were passed on to designated district officials, who could send a follow-up message back to the vendor.

### 4.1.2 Top-Down Treatment Bundle

The second treatment arm was designed to improve district governments' capacity to collect taxes, to reduce the leakage of revenues as they were transferred to the district governments, and to improve the incentives faced by tax collectors and market masters to collect taxes honestly. Markets assigned to receive the top-down treatment received the following four components:

**Step 1: Roll Out Mobile Money Revenue Transfer System.** Prior to our study, tax collectors collected fees in cash from vendors and gave the cash to the market manager, who then remitted the cash to the district in person a few times a month. As this system provided many opportunities for corruption and obscures revenue tracking, treatment markets converted to a system in which market managers deposit market fee revenue daily into the district bank account via a mobile money agent.

**Step 2: Provide Accurate and Reliable Market Vendor Counts.** At baseline, district governments had almost no information about the revenue potential at each market, which is primarily a function of market size. To address this, trained vendor counters visited each treatment market four times a month during the intervention period.

**Step 3: Generate Market Revenue Targets.**

The vendor counts from Step 2 were fed into a revenue target calculator that created monthly targets for each market based on seasonality and the previous month's revenues. These targets were then communicated to market managers and tax collectors. For tax collectors, this provides a check against corruption and serves as an incentive for better performance.

**Step 4: Introduce Incentives for Tax Collectors.**

Finally, treatment markets received a tax collector incentive system using the revenue targets created in Step 3. If a market met its monthly revenue target, district government presented the market with valuable goods, typically wheelbarrows and bicycles, that facilitated market management.[4]

## 4.2 Hypotheses

Each treatment bundle was designed to address community-level barriers to tax payment. Following Section 2, we expected each arm of our experiment to increase the fraction of

[4] This component originally included individual incentives for each tax collector. These were eliminated after the first month due to district government concerns. In practice, incentives were frequently delivered late, and were not always given according to performance criteria. These complications may have weakened the top-down treatment.

vendors who pay market fees. We also expected each treatment to increase the revenue district governments receive from each market. Our main hypotheses are therefore that:

**H1:** Each treatment will increase the percentage of taxpayers who pay their fees
**H2:** Each treatment will increase the revenue per market that the government receives.[5]

Our final treatment group gets both the top-down and bottom-up intervention bundles. We expected this combined treatment to be more effective than either approach alone, theorizing that better enforcement could actually support quasi-voluntary compliance. This comes out of arguments that using audits and penalties to compel those who don't pay voluntarily can bolster tax morale among those who do (Coleman, 2007).

**H3:** The two treatment arms will have the largest effect in combination.

However, based on qualitative evidence we gathered during the intervention period, our discussion section, below, considers alternative possibilities, such as the possibility that enforcement could "crowd out" intrinsic motivation among those with high tax morale if enforcement lowers the perceived legitimacy of government (Dwenger et al., 2016).

We also pre-specified hypotheses regarding the causal mechanisms and downstream effects we expected from each treatment arm.[6] For the bottom-up treatment, which focuses on tax bargaining, we expected it to:

**H4:** Increase taxpayers' trust in government
**H5:** Increase taxpayers' satisfaction with the government
**H6:** Increase taxpayers' satisfaction with the level of market services
**H7:** Increase taxpayers' tax morale

In bottom-up markets, we also expect to see downstream effects on political engagement, as this bundle of interventions aims to empower vendors to advocate on their own behalf. As

[5] The focus of this paper is taxpayer behavior and outcomes, so we refrain from extensively discussing and testing this hypothesis. The full pre-specified set of analyses for this hypothesis is available in Appendix E.

[6] We have reordered the hypotheses from the order they were in the PAP. H8 below was considered an "indirect effect" and was listed as H11. However, this was the only indirect effect hypothesis and is only a hypothesis about the bottom-up treatment, and so we decided that it fits more naturally with the other bottom-up hypotheses.

such, we hypothesize that in bottom-up markets:

**H8:** Vendors will become more politically engaged

For the top-down treatment, which requires the state to invest in more monitoring and bureaucratic capacity, we expect it to affect compliance through the following causal mechanisms:

**H9:** Increase enforcement of the tax
**H10:** Decrease corruption
**H11:** Increase tax collector effort

## 5 Empirical Strategy

To measure our outcomes, we collected survey data from market vendors and tax collectors.[7] In each market we surveyed 100 vendors at baseline and endline. Vendors were chosen via a modified random walk (see Appendix C.1.1 for details), and different individuals were sampled at baseline and endline. Of the 100 vendors, 80 received a 15-minute survey measuring tax compliance and demographics. A randomly-chosen 20 vendors received a longer, 1-hour survey that included additional mechanism and treatment compliance questions. The larger per-market sample for the tax compliance questions allows more precise market-level estimates for those outcomes. Markets were visited on their main market day when the largest number of vendors were present. Vendors received a small airtime voucher for completing the survey. Total sample size is 12,389 at baseline and 12,370 at endline.

Enumerators also surveyed each market's tax collectors. The survey covered job details, perceptions of vendor compliance and relations, and knowledge of intervention components. Our baseline sample has 302 tax collector surveys; at endline this is 264. On average 2-3 tax collectors were interviewed in each market.

[7] See Appendix C for an in-depth explanation of data sources. All surveys were implemented by Innovations for Poverty Action in Malawi. See Appendix D for survey descriptive statistics.

For both surveys, we describe our main measures in the following sections, and report additional measurement details in Appendix C.3. Our pre-analysis plan also specified that we would analyze monthly tax revenue information for each market in our sample. However, these data, provided by low-capacity district governments, proved to be of poor quality. We include this analysis, along with a discussion of data quality, in Appendix E.

## 5.1 Empirical Models

For our analysis the main independent variables are indicators for whether a market received the top-down treatment only, the bottom-up treatment only, or both treatments. While our pre-analysis plan specified analyzing the "BOTH" condition using interaction effects, qualitative feedback from the intervention period suggests that it operates more as a distinct treatment experience, and less as the combination of the two individual treatments. We therefore analyze it as a separate third treatment, rather than as an interaction. Appendix M.1.3 reports the pre-specified interaction analysis.

Because our sample is a repeated cross-section, not a panel, all individual-level regressions are performed only using the results of the endline survey. All individual-level regressions take the following form:

$$Y_{ijkl} = \beta_0 + \beta_1 * BU_j + \beta_2 * TD_j + \beta_3 * BOTH_j + \beta_k * ENUM_k + \beta_l * Block_l + \epsilon_{ijkl}$$

Where $Y_{ijkl}$ represents an outcome measure for vendor $i$ in market $j$ in block $l$, interviewed by enumerator $k$, measured at endline. $TD_j$, $BU_j$, and $BOTH_j$ are indicators that are 1 if market $j$ was in that treatment group and 0 if not. As discussed above, the BOTH treatment ended up being a distinct treatment from simply the combination of the BU and TD treatments, so we include separate dummies for the three different treatment groups. We include enumerator fixed effects ($ENUM_k$) because enumerator skill and general behavior can impact respondents' answers. We include block fixed effects ($Block_l$) to control for unobservable differences between the blocks. Because treatment was assigned at the market level, we cluster standard errors at that level.

In addition to the individual-level analysis, we perform market-level regression analyses.[8] First, we perform a simpler version of the individual-level analysis, with the endline outcomes averaged to the market level. These regressions take the following form:

$$Y_{jl} = \beta_0 + \beta_1 * BU_j + \beta_2 * TD_j + \beta_3 * BOTH_j + \beta_l * Block_l + \epsilon_{jl}$$

$Y_{jl}$ represents the average endline outcome for market $j$ in block $l$. As above, $TD_j$ and $BOTH_j$ are indicators that are 1 if market $j$ was in that treatment group and 0 otherwise we once again include block fixed effects to control for differences between the districts.

Second, we estimate a difference-in-differences (DID) model. The actual model is the same as the market-level endline model described above, but $Y_{jl}$ is now the difference in the average endline outcome between endline and baseline for market $j$, i.e. $Y_{jl} = Y_{jl(Endline)} - Y_{jl(Baseline)}$. This is equivalent to the typical one-time period DID estimator and is more easily interpretable. In this model, $\beta_1$, $\beta_2$, and $\beta_3$ represent changes in the changes from Baseline to Endline in the BU, TD, and BOTH groups compared to the control group.

The coefficient estimates for all the treatment indicators represent intent-to-treat (ITT) estimates, as implementation was inconsistent and ITT estimates represent the most conservative estimates.

[8] We do this for measures that were included on the long and short survey versions, meaning we have 100 observations per market to use in the average. For outcomes only included on the long survey (20 respondents per market) we rely only on individual-level measures, as the market-level measures are too noisy.

# 6 Results

## 6.1 Treatment Effects: Tax Compliance

Hypothesis 1 predicted that each treatment would improve individual-level tax compliance, and Hypothesis 3 predicted that the BOTH group would have the largest treatment effects. Our primary measure of tax compliance is the same as that used to do the block randomization: whether the vendor can produce a tax receipt from within the past 7 days. We pre-specified this as our main measure because it is verifiable and so less subject to response bias than self-reported measures. Our analysis uses both an individual- version of this variable, and an aggregated market-level version. Table 2 reports the results for this analysis.

Supporting Hypothesis 1, the verified receipt measure provides evidence for an increase in tax compliance in the two treatment arms, although the results are strongest for the BU treatment. Compared to control group vendors, BU (TD) vendors were 10.1 (7.4) percentage points more likely to be able to provide a recent receipt; these differences are statistically significant. Contrary to Hypothesis 3, we do not see larger effects in the BOTH group: there, tax compliance increased by roughly six percentage points, although this effect falls short of statistical significance (p=0.07). In the DID model, only the coefficient for the BU treatment arm is significant. However, as the the DID estimates rely on market-level averages they are inherently more conservative.

We view this as the most compelling evidence for the intervention's positive impact on tax compliance, as it required vendors to show enumerators a physical receipt. It is important to note that this measure of tax compliance is most likely a conservative one in and of itself, as it requires individuals to retain receipts. This is revealed in mean compliance according to each measure: 32.6% for the receipt measure, compared to 78.9% for the self-reported compliance measure.

Appendix Table M1 in Appendix M.1.1 reports two alternative measures of tax compliance, including a measure of self-reported compliance, and a measure of whether a respondent perceives others as complying. These measures are not moved by the treatments. As the self-reported measures suffer from social desirability bias, especially the "own compliance" measure, we put more weight on our verified receipt measure.

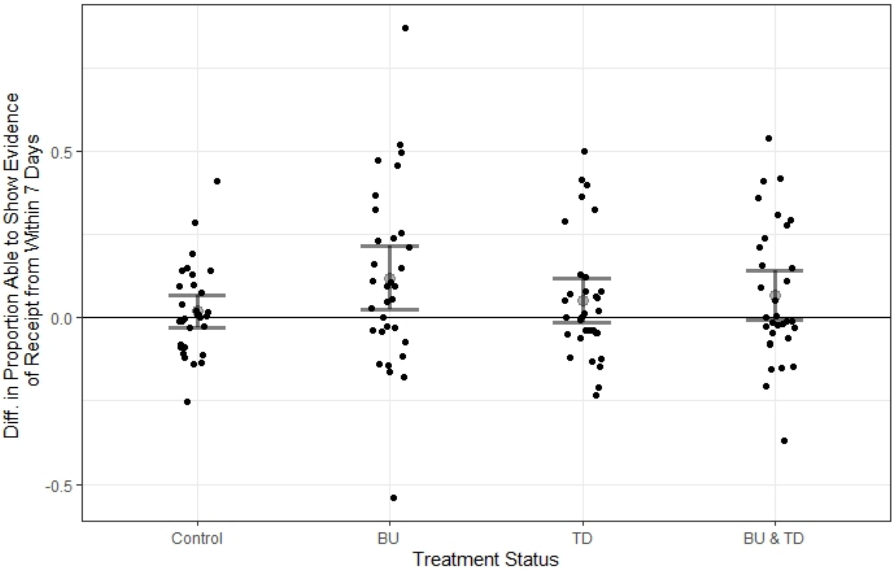
Table 2: Hypothesis 1 Results Table - Individual-Level DIM and Market-Level DID

| | Evidence of Receipt from Past 7 Days | |
|---|---|---|
| | Individual DIM | Market DID |
| BU | 0.101** | 0.100* |
| | (0.031) | (0.045) |
| TD | 0.074* | 0.034 |
| | (0.030) | (0.045) |
| Both | 0.057 | 0.050 |
| | (0.031) | (0.045) |
| Observations | 12,365 | 128 |
| Adjusted $R^2$ | 0.268 | 0.211 |

Notes: *p<0.05; **p<0.01; ***p<0.001
Individual-level model includes enumerator and block fixed-effects
Individual-level model has SEs clustered on market.
Market-level model includes block fixed-effects.

To further explore the results, Figure 1 shows the difference in the market-level averages for the receipt measure for each of the four possible treatment assignments. The solid black points represent differences for each of the markets. The larger translucent grey points indicate the treatment group mean, weighted by respondents per market, with lines indicating the 95 percent confidence interval, calculated using cluster-adjusted standard errors. As in the regression results, Figure 1 shows the dramatic change from baseline to endline in the proportion of individuals able to present a recent receipt for fee payment for all treatment

Figure 1: Evidence of Recent Receipt: Difference between Baseline and Endline



groups, with the biggest change happening in the BU treatment group.[9]

One concern is that the receipt measure could reflect vendors' ability to get a receipt, rather than higher compliance. To test this, Appendix Table M7 reports the results of an analysis using a survey question in which we asked, on a 5-point scale, how often "you pay the fee but do not get a receipt." To make the results more comparable to the receipt measure, we dichotomized this outcome: vendors who said "it happens", "it happens a lot," or "always" were coded as 1, otherwise they were coded as 0. While we do find significant decreases in "pay but no receipt" in the BU and BOTH groups, the point estimate is too small (5.9 percentage points) to account for the 10 percentage point increase we see in the percent of vendors who have a recent tax receipt in Table 2.

Together, the analysis suggests an increase in vendor tax compliance in the BU and TD groups, although the evidence is strongest in the BU treatment group. Interestingly, and contrary to Hypothesis 3, treatment effects are weakest in the markets that received both treatments; we discuss this pattern further below.

[9] Appendix Figure M1 reports an alternative visualization that shows how each individual market's compliance level changed between baseline and endline.

## 6.2 Causal Mechanism And Downstream Effects: Bottom-up Treatment

Our pre-analysis plan posited that, if the treatments work, it would be due to specific causal mechanisms, as specified in Section 4.2. Hypotheses H4-H7 argued that the bottom-up treatments should increase taxpayers' trust, satisfaction with government, satisfaction with services, and tax morale. Hypothesis 8 argued that vendors' political engagement would also be higher. These outcomes, while critical for interpreting the main treatment effects, are also important outcomes for state capacity and development more broadly. If the proposed mechanisms are moved by the treatment, it establishes the base for quasi-voluntary compliance and trust in and willingness to engage with the state. Higher trust could then support the expansion of state capacity in other ways.

To test Hypotheses 4-7, we use outcomes from the market vendors' survey. Trust is measured by two questions, which asked respondents about how trustworthy (H4) their district government and ward councilors are on a four-point agreement scale. To measure satisfaction with government (H5), we asked respondents how strongly they agreed with how their district government manages public funds, transparency in how it uses funds, and transparency in how it collects market fees. To measure satisfaction with services (H6), we use average self-reported satisfaction with five common market services, measured on a 4-point scale. We test robustness using a question that asked respondents how satisfied they were with "the developments in this market provided by the district government." Finally, to measure tax morale (H7), we used two questions. The first asked respondents whether they agreed or disagreed (on a 4-point scale) with the statement "Paying taxes is a duty of all citizens, even when you do not approve of how elected officials spend money." The second asked respondents whether they thought that vendors should pay tax even if they disagree with local government, or only when they agreed with the local government. For analysis, we turned this into a binary outcome that was 1 if a respondent stated that vendors should pay even if they disagree, and 0 otherwise. The second tax morale question, and the services satisfaction question, were asked to all respondents; the other measures were only asked to the 20 respondents per market who took the long version of the survey. As pre-specified, we only use individual-level DIM estimation for these outcomes.

Table 3 reports the results for the trust and government satisfaction measures (H4-5). To aid interpretation, our discussion here translates the raw coefficients to the percent increase, relative to the control group. As predicted, we find that the BU treatment significantly increased trust in the district government (6.7% increase over control once converted to a percent increase [10]) and ward councilor (6.6% increase over control). However, this increase appears to be independent of general perceptions of government performance: there is no accompanying change in perceptions that the district manages fund well, or manages revenue and spending transparently.

However, we do see significantly higher satisfaction with market services more specifically. Columns 1 and 2 of Table 4 show that vendors exposed to the BU treatment were more satisfied with market services in general than those in the control group (14.7% increase). This result is driven by a large increase in satisfaction with access to clean water (33.2% increase for the BU treatment group; 16.0% increase for the Both treatment group), likely because boreholes were the chosen construction project in 43 of 64 BU treatment markets. These gains in satisfaction are not driven by perceptions that the government in spending more overall on services (Column 3).

Our results on tax morale are mixed. While we do find a significant increase in the fraction of BU vendors who report paying taxes is a duty (Column 4 of Table 4), the substantive effect is small (2% increase over control), and we see no movement in our second measure, which asked whether vendors should pay taxes always, or only when they agree with district government. This is consistent with a world in which citizens are conditional compliers, and the treatment increased satisfaction and willingness to pay taxes.

[10] See Table G1 in Appendix G for the substantive impact for the non-dichotomous mechanism outcomes as percent increase over the control group average.

Overall, we interpret these results as supporting many of the proposed causal mechanisms for the BU treatment: in interacting more with district government officials and experiencing more responsiveness surrounding revenue collection and service provision, vendors feel more trusting, satisfied, and duty-bound vis-a-vis their government. We note that those vendors who did not experience the BU treatment did not display similar spikes in trust, satisfaction, or tax morale, and that the results are weaker for the BOTH treatment group. Although vendors in the different treatment groups did not view district government as doing a better job at managing funds and being more transparent, it is possible that these questions were either too technical for vendors, or that other factors, such as support for the party in power, are better predictors for answers to these questions.

Table 3: Bottom-Up Causal Mechanism Outcomes: H4 - H5. Individual-level models include enumerator and block fixed-effects. Individual-level models have SEs clustered on market.All outcomes are on a 4-point scale.

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Trust in Local Gov. | Trust in Ward Counc. | Dist. Manages Funds Well | Dist. Transp. Spending | Dist. Transp. Tax Collection |
| | *OLS* | *OLS* | *OLS* | *OLS* | *OLS* |
| BU | $0.176^{**}$ | $0.168^{*}$ | −0.087 | −0.076 | −0.058 |
| | (0.063) | (0.070) | (0.058) | (0.067) | (0.063) |
| TD | 0.001 | −0.117 | −0.008 | −0.025 | −0.026 |
| | (0.068) | (0.062) | (0.058) | (0.069) | (0.054) |
| Both | $0.142^{*}$ | 0.103 | −0.061 | −0.039 | −0.071 |
| | (0.059) | (0.066) | (0.056) | (0.056) | (0.050) |
| Observations | 2,509 | 2,447 | 2,521 | 2,518 | 2,510 |
| Adjusted $R^2$ | 0.182 | 0.112 | 0.332 | 0.373 | 0.381 |

$^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001

## 6.3 Political Engagement

In addition to the causal mechanisms, we also predicted that the BU treatment would increase political engagement among vendors (H8). To test this, we examine four outcomes. The first

Table 4: Bottom-Up Causal Mechanism Outcomes: H6 - H7. Individual-level models include enumerator and block fixed-effects. Individual-level models have SEs clustered on market. Outcomes 1, 2, and 4 are on 4-point scale. Outcome 5 is dichotomous. Outcome 3 is a number out of 1000.

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Services Satisfaction | Satisfaction with Water Access | Percep. of Spending on Services | Paying Tax as Duty | Tax Morale |
| | *OLS* | *OLS* | *OLS* | *OLS* | *OLS* |
| BU | $0.293^{**}$ | $0.654^{***}$ | 27.576 | $0.072^{*}$ | 0.002 |
| | (0.095) | (0.160) | (15.538) | (0.035) | (0.012) |
| TD | 0.104 | 0.161 | 7.046 | 0.041 | 0.007 |
| | (0.087) | (0.129) | (15.020) | (0.030) | (0.011) |
| Both | 0.173 | $0.315^{*}$ | 0.515 | 0.044 | 0.022 |
| | (0.092) | (0.148) | (14.205) | (0.033) | (0.013) |
| Observations | 12,365 | 2,517 | 2,411 | 2,531 | 12,355 |
| Adjusted $R^2$ | 0.161 | 0.140 | 0.290 | 0.111 | 0.082 |
| Notes: | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001 | | | | |

measures whether respondents report planning to vote in the next election. Measures 2 and 3 are based on whether respondents are willing to sign a hypothetical petition to the District Finance Committee telling them that vendors want the District to improve funding for markets to assess how the treatments potentially increase bottom-up pressures. Respondents were first asked whether they would be willing to sign the petition anonymously, and then whether they would be willing to sign with their names. Our last measure is a behavioral outcome. We asked respondents to send an SMS agreeing with two different statements concerning the government: 1) "We would like to demand a reduction in the over-reach of government power in collecting revenue in this district," and 2) "We would like to demand an increase in government effectiveness in spending government revenue collected in this district."[11] Respondents also had the chance to write a longer message, which they were informed would be passed on to the district government along with the level of agreement.

[11] The order of the two statements was randomized.

Names and numbers were not provided to the government.

In line with our expectations, we find evidence that vendors became more politically engaged as a result of the interventions in markets that received the bottom-up treatment. There is no effect on anticipated voter turnout, although this may be because of ceiling effects – roughly 85% of vendors in all treatment groups reported intending to vote. This may have been because the next elections were only 5 months away; it is also consistent with social desirability bias leading to overreporting of vote intention.

We see stronger effects for willingness to sign a petition in the BU and BOTH groups, both anonymously and by name. In addition, vendors in BU and Both markets were 5-6 percentage points more likely to send an SMS message agreeing with the statement "We would like to demand a reduction in the over-reach of government power in collecting revenue in this district" than vendors in control markets. Vendors in "Both" markets were also roughly 5 percentage points more likely to send an SMS message agreeing with the statement "We would like to demand an increase in government effectiveness in spending government revenue collected in this district."

The finding that citizens receiving the BU treatment demand a reduction in over-reach by the local government seems counter-intuitive given our findings in the mechanism tests that there is *increased* trust in the local government. If citizens trust the government, why would they want to limit the ability of the local government to collect taxes? There are several possible explanations for this dynamic. First, it is possible that citizens view the two as separate. They may have gained increased trust in the government overall because of the bottom-up treatment components, but may still be hesitant to trust the enforcement arm of the government. Second, it is possible that because of their increased trust, they feel that the potentially harsh methods that local governments can employ to collect revenue should be unnecessary. However, in the next section we show that the BU treatments did not increase the perceived coerciveness of tax collection, alleviating the concern that our interventions led to that the effects are not due to overly-coercive government practices. The finding that citizens receiving the BU treatment demand a reduction in over-reach by the local government seems counter-intuitive given our findings in the mechanism tests that there is *increased* trust in the local government. If citizens trust the government, why would they want to limit the ability of the local government to collect taxes? There are several possible explanations for this dynamic. First, it is possible that citizens view the two as separate. They may have gained increased trust in the government overall because of the bottom-up treatment components, but may still be hesitant to trust the enforcement arm of the government. Second, it is possible that because of their increased trust, they feel that the potentially harsh methods that local governments can employ to collect revenue should be unnecessary. However, in the next section we show that the BU treatments did not increase the perceived coerciveness of tax collection, alleviating the concern that our interventions led to overly-coercive government practices.

Finally, it is possible that we are seeing two distinct vendor reactions to the BU treatments. One set of vendors experiences stronger trust in government and a sense that taxpaying is a duty, while another set of vendors is unhappy with government involvement in markets. We see some descriptive evidence of this - vendors from BU and Both treatment markets who agreed to send the message had lower levels of trust in the local government than vendors who did not agree to send the message (a decrease of 7.6%). In addition, refitting Model 4 in Table 5 and including an interaction between the treatment variables and trust in government reveals that individuals in BU markets who had indicated that they considered the government "Very Trustworthy" were 3% *less* likely to agree with the statement about over-reach.[12]

The TD treatments did not seek to empower vendors in the same way as the BU treatments. As expected, there are consistent null effects on the political engagement outcomes in the TD markets. It is encouraging that the stronger enforcement in the TD markets did not

[12] See Appendix H for the associated models.

Table 5: Political Engagement Outcomes

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Vote | Petition Anon. | Petition w. Name | Agree St. 1 | Agree St. 2 |
| BU | 0.001 | $0.057^{**}$ | $0.069^{*}$ | $0.058^{**}$ | 0.031 |
| | (0.017) | (0.021) | (0.027) | (0.021) | (0.021) |
| TD | −0.003 | −0.011 | −0.007 | −0.004 | −0.011 |
| | (0.017) | (0.023) | (0.028) | (0.020) | (0.021) |
| Both | −0.021 | $0.066^{**}$ | $0.081^{**}$ | $0.054^{**}$ | $0.048^{*}$ |
| | (0.018) | (0.023) | (0.030) | (0.019) | (0.021) |
| Observations | 2,527 | 2,514 | 2,514 | 2,531 | 2,531 |
| Adjusted $R^2$ | 0.030 | 0.221 | 0.237 | 0.318 | 0.359 |

Notes: $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.
All models linear probability models.

lead to demands to reduce tax collection (which is what a statistically significant effect on agreement with the statement about over-reach would suggest).

## 6.4 Causal Mechanism And Downstream Effects: Top-Down Treatment

We expected the top-down treatment to increase perceived and actual tax enforcement (H8); decrease corruption (H9); and increase tax collector effort (H10). To measure these causal mechanisms, we use questions from the market vendors and tax collector surveys. From the market vendors' survey, we include a number of measures. To measure perceived coercive pressure to pay the tax (H8), we use three questions that measure vendors' perceptions that they could individually or collectively avoid market fee payment, plus a question that asks whether respondents agree or disagree on a 4-point scale that "I pay market fees because

I'll get in trouble if I don't." To measure citizen perceptions of corruption in fee collection (H9), we use a question measuring the perceived fraction of fees that never reach the district government.[13]

We also use several measures from the tax collectors' survey. First, we use a a list experiment to estimate bribe-taking as an additional measure of corruption (H9); tax collectors were given a list events, and asked how many had happened to them in the previous week. In the treatment group, one item was "A vendor in this market gave you money so that vendor did not have to pay the market fee." To measure tax collector effort (H10), we use questions that measure the number of hours worked per day by each tax collector, and the average number of vendors visited each day.

Tables 6 and 7 report results from the market vendor and tax collector surveys, respectively. We find little evidence of increased coercive tax enforcement; while vendors in the top-down group are slightly more likely to report that they pay taxes due to the consequences of non-payment, this effect is substantively small (1.5% increase over control), and there is no effect on vendors' beliefs about their ability to refuse to pay the fee, either alone or together. In all conditions, large majorities of vendors disagreed with the assertion that noncompliance was possible.

We find mixed evidence that the treatments decreased corruption. There is no evidence that vendors perceive lower levels of tax collector corruption in TD markets, as measured by the perceived fraction of tax revenues that actually reach the district. We do see that vendors in BOTH treatment markets reported that they thought more money flowed to the government than in control markets, but the effect size is substantively small (3.6% increase over control).

To further measure the treatments' effects on corruption, we included a list experiment on

[13] Our PAP also specified a measure estimating individual compliance using our survey measures, then comparing this to the actual revenues reported by the district government. The revenue data quality issues discussion in Appendix E led us to drop this measure.

the tax collector survey. The control group was asked how many of four innocuous activities had happened to them in the past week; the treatment group's list also included a fifth item that asked about bribe-taking. Across all treatment groups, the list experiment estimates that 14% of tax collectors report accepting money from a vendor seeking to avoid paying the market fee. However, this varies greatly by treatment group: we find corruption estimates of about 0% in the TD group; 4% in the BOTH group; 18% in BU markets; and 36% in control markets. While due (at least in part) to the small sample size these differences are not statistically significant, the results are consistent with a world in which bribe-taking is lower in markets that got the TD treatments.

Table 6: Top-Down Causal Mechanisms Outcomes, Vendor Survey. Individual-level models include enumerator and block fixed-effects, and cluster SEs by market. Models 1, 2, and 3 are on a 4-point scale. Model 4 is 0-1000.

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Ind'l Evasion Possible | Group Evasion Possible | Pay Because Consequences | Money Flowing to Gov't |
| BU | −0.052 | 0.016 | 0.036 | 18.371 |
| | (0.057) | (0.059) | (0.027) | (13.759) |
| TD | −0.055 | 0.059 | $0.056^{*}$ | −2.808 |
| | (0.052) | (0.057) | (0.025) | (12.106) |
| Both | −0.046 | −0.058 | 0.041 | $26.126^{*}$ |
| | (0.058) | (0.060) | (0.028) | (11.219) |
| Observations | 2,514 | 2,524 | 2,518 | 2,463 |
| Adjusted $R^2$ | 0.123 | 0.144 | 0.308 | 0.257 |
| Notes: | | | | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001 |

We find stronger evidence that the TD treatment increased tax collector effort. While the treatments did not increase the number of vendors tax collectors report visiting, tax collectors do report spending significantly more time in TD markets at endline. This increase represents a 12.5% increase over the time spent by tax collectors working in control markets. This suggests that that tax collectors in TD markets are spending more time with the vendors they do visit, in line with other studies of tax collector incentives (see e.g. Khan, Khwaja and Olken (2016)). This could be because tax collectors in TD markets felt more pressure from market management due to incentives and more scrutiny from the district government.

Table 7: Top-Down Causal Mechanisms Outcomes, Tax Collector Survey

| | *Dependent variable:* | |
|---|---|---|
| | Hours Working in Market A Day | Vendors Visited Per Day |
| BU | 0.304 | 59.100 |
| | (0.581) | (64.785) |
| TD | 1.154* | 85.199 |
| | (0.494) | (59.896) |
| Both | 0.609 | 162.246 |
| | (0.562) | (116.823) |
| Observations | 264 | 261 |
| Adjusted $R^2$ | 0.367 | 0.256 |

Notes: *p<0.05; **p<0.01; ***p<0.001
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

## 7 Discussion

The results show that both the top-down and bottom-up treatments affected state capacity. The bottom-up treatment significantly increased tax compliance. The top-down and cross-cutting treatments had a smaller, less robust impact on tax compliance. We also see movement in some of our expected causal mechanisms and downstream effects. These mechanism tests also produce results that are important in their own right: increasing government trust and satisfaction is important for a wide range of governance outcomes, and increasing tax collector effort is likewise important.

In contrast, the markets that received both the top-down and bottom-up treatments saw smaller treatment effects, both for the main outcomes and the associated mechanisms. There are several potential reasons for this. First, given the low level of state capacity in Malawi, it is possible carrying out both treatments was too demanding, resulting in the treatment backfiring due to ineffective execution. It is also possible that one treatment crowded out the other. For example, it is possible that while the BU treatments caused tax morale to increase, adding in the TD treatments, which can lead to higher pressure on vendors, dampened these effects by undermining vendors' sense that they were paying for their own development. Finally, it is possible that delays in the construction component, in conjunction with the increased TD pressure, meant that vendors faced increased enforcement after they had been promised new public goods, but before those public goods were actually delivered.

Interestingly, markets that received both treatments saw similar or stronger effects than BU only markets for the political engagement outcomes. This suggests that although the BU treatment bundle did not have the desired effect on tax compliance, it did still empower vendors, increasing their willingness to advocate on their own behalf.

The rest of this section describes robustness and potential threats to inference.

## 7.1 Robustness Checks

The results for self-reported and group-perceived tax compliance and the receipt measure are robust to alternative specifications, including market-level endline difference-in-means, market-level difference-in-differences, individual-level quasi-difference-in-differences[14], and individual-level endline difference-in-means controlling for the baseline market-level average for the outcome variable (see App. M.1.2). We also analyzed the main outcomes using an interaction between top down and bottom-up treatment arms (in other words, analyzing the

[14] "Quasi" because we do not have panel data; instead we assume that baseline and endline respondents are drawn from the same population. These models are much noisier than panel difference-in-differences.

experiment as a factorial design); these models show the same results. The results are also robust to different formulations of the outcome variables, including retaining observations with nonsensical values for the self-reported and group-perceived tax compliance measures (see App. M.1.4), and widening the receipt window to ten days (see Table M6 in App. M.1.5). Our mechanism outcomes are similarly robust to different modeling approaches, including individual level quasi-difference-in-differences and endline difference-in-means controlling for the baseline market-level DV (see App. M.2).

We considered several possible heterogeneous effects, but generally find limited evidence that effects vary across sub-groups (see Appendix Section L). Specifically, we pre-specified exploring heterogeneous effects by vendor gender, vendor type (selling goods versus services), frequency of vending (daily versus not daily), and vendor wealth, but do not find any evidence of heterogeneous effects based on these vendor-level variables.

In addition, based on our qualitative data regarding treatment implementation, we opted to explore heterogeneous effects by market size and by market propensity for collective action. We operationalize market size using the number of vendors as counted by the implementing partner during the pre-treatment scoping phase on a market day, and find no heterogenous effects based on this variable. We operationalize collective action propensity by calculating the market-level mean of a question on the vendor survey inquiring whether the market would work together to solve a problem. We find that the likelihood of producing a receipt does not vary with collective action propensity, but that self-reported and group-perceived tax compliance effects in the bottom-up group are stronger as collective action propensity increases.

Because many vendors sell in multiple markets, and because markets can operate in close proximity to one another, we also conducted spillover analysis using two approaches: an inverse probability weighting (IPW) approach and a treatment externalities approach based on Miguel and Kremer (2004). Full description of and results from this analysis are in

Appendix I. Both approaches find treatment effects robust to most specifications. Cases where our results are weaker (a 10 km radius in the IPW approach, and one classification coding in the treatment externalities approach) could be driven by the drop in observations. They are also consistent with a world in which our treatment effects are primarily driven by large markets in dense urban areas.

Finally, we carried out pre-specified multiple hypothesis correction, to assess how robust our results are to false positive conclusions. As laid out in our pre-analysis plan, we performed corrections per hypothesis, collecting all tests for all outcomes for each hypothesis and using the Holm procedure to see which $p$-values survive.[15] Table N1 in Appendix N shows the original and corrected $p$-values. Multiple hypothesis correction solidifies our belief in the effect in the bottom-up treatment group, as all except two of the significant results for the BU term survive – the market-level recent receipt result, which was a hard test to begin with due to the the relatively small sample size at the market-level, and the petition with name outcome for H8, although the corrected $p$-value is only just above the 0.05 cutoff at 0.060.

## 8 Conclusion

This paper presents some of the first evidence on how best to improve state capacity and tax compliance in a developing country setting. We implemented a set of complex, multi-pronged interventions that were designed to provide a strong test of two common approaches to improving taxation. The first, bottom-up approach focused on improving citizen-state relations, jump starting tax bargaining, and providing a way out of the low-services, low-compliance

[15] For example, for H1, we considered together the tests for difference from zero for the BU, TD, and Both coefficients for the three main outcomes – self-reported tax compliance, group-perceived tax compliance, and the recent receipt measure – at the individual- and market-level. In other words, we corrected for doing eighteen tests for H1. For the BU hypotheses and TD hypotheses, as these were about the BU and TD treatments, respectively, we only corrected for the BU and Both coefficient tests and the TD and Both coefficient tests, respectively.

tax equilibrium. The second top-down arm focused on enforcement and government capacity to monitor and collect taxes efficiently. Critically, we focused on interventions that targeted communities of taxpayers, rather than individuals; this more closely matches the theoretical predictions of past work on taxation.

The Bottom-Up intervention successfully increased tax compliance and improved citizen-government relations and citizen involvement. This suggests that such approaches may be critical for creating sustained improvements in taxation and state compliance in developing countries. In contrast, we find smaller effects on tax compliance in markets that received the top-down treatment arm, or the combined treatment arm. The latter result suggests that, especially in low-capacity states, trying to implement too many changes at once may backfire.

We also find encouraging evidence that our interventions improved markets in other ways. In the bottom-up treatment group in particular, the intervention increased trust in government and satisfaction with services. This would be a valuable outcome even absent an effect on tax compliance. In addition, the bottom-up interventions seemed to increase political engagement among vendors; this is another important intermediate outcome with positive implications for citizen-state relations.

By the nature of the design, it is difficult to determine which intervention components are driving these effects. Future work will be needed to determine the most effective treatments. More work is also needed on the political feasibility of each approach. Both treatment arms ran into a lack of political will to implement the experiment as originally agreed upon: local officials worried about giving detailed spending information to citizens, and about the ways in which incentive schemes for tax collectors could backfire. However, that relatively new governments were able to implement such an ambitious set of treatments points to the potential for this kind of approach even in low-capacity settings.

# References

Agrawal, Arun, Ashwini Chhatre and Elisabeth Gerber. 2015. "Motivational Crowding in Sustainable Development Interventions." *American Political Science Review* 109(3):470–487.

Allingham, Michael G and Agnar Sandmo. 1972. "Income tax evasion: A theoretical analysis." *Journal of public economics* 1(3-4):323–338.

Alm, James, Betty R Jackson and Michael McKee. 1992. "Estimating the determinants of taxpayer compliance with experimental data." *National Tax Journal* pp. 107–114.

Alm, James, Jorge Martinez-Vazque and Benno Torgler. 2006. "Russian attitudes toward paying taxes–before, during, and after the transition." *International Journal of Social Economics* .

Andreoni, James, Brian Erard and Jonathan Feinstein. 1998. "Tax compliance." *Journal of economic literature* 36(2):818–860.

Baskaran, Thushyanthan and Arne Bigsten. 2013. "Fiscal Capacity and the Quality of Government in Sub-Saharan Africa." *World Development* 45:92–107.

Bates, Robert H. and Da-Hsiang Donald Lien. 1985. "A Note on Taxation, Development, and Representative Government." *Politics and Society* 14(1):53–70.

Brollo, Fernanda, Tommaso Nannicini, Roberto Perotti and Guido Tabellini. 2013. "The Political Resource Curse." *American Economic Review* 103(5):1759–96.

Castro, Lucio and Carlos Scartascini. 2015. "Tax compliance and enforcement in the pampas evidence from a field experiment." *Journal of Economic Behavior & Organization* 116:65–82.

Coleman, Stephen. 1996. "The Minnesota income tax compliance experiment: State tax results.".

Coleman, Stephen. 2007. "The Minnesota income tax compliance experiment: replication of the social norms experiment." *Available at SSRN 1393292* .

De la Cuesta, Brandon, Lucy Martin, Helen V Milner and Daniel L Nielson. 2022. "Owning it: Accountability and citizens' ownership over oil, aid, and taxes." *The Journal of Politics* 84(1):304–320.

Del Carpio, Lucia. 2013. "Are the Neighbors Cheating? Evidence from a Social Norm Experiment on Property Taxes in Peru Job Market Paper.".

Dwenger, Nadja, Henrik Kleven, Imran Rasul and Johannes Rincke. 2016. "Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany." *American Economic Journal: Economic Policy* 8(3):203–32.

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

Fjeldstad, Odd-Helge and Ole Therkildsen. 2008. "Mass taxation and state-society relations in East Africa." *Taxation and State Building in Developing Countries* .

Frey, Bruno S. and Reto Jegen. 2001. "Motivation Crowding Theory." *Journal of Economic Surveys* 15(5):589–611.

Hallsworth, Michael, John A List, Robert D Metcalfe and Ivo Vlaev. 2017. "The behavioralist as tax collector: Using natural field experiments to enhance tax compliance." *Journal of public economics* 148:14–31.

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

Kleven, Henrik Jacobsen, Martin B Knudsen, Claus Thustrup Kreiner, Søren Pedersen and

Emmanuel Saez. 2011. "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark." *Econometrica* 79(3):651–692.

Levi, Margaret. 1989. *Of Rule and Revenue.* University of California Press.

Levi, Margaret, Audrey Sacks and Tom Tyler. 2009. "Conceptualizing legitimacy, measuring legitimating beliefs." *American behavioral scientist* 53(3):354–375.

Mascagni, Giulia, Andualem Mengistu and Firew B Boldeyes. 2018. "Can ICTs increase tax? Experimental evidence from Ethiopia." *International Centre for Tax and Development, Working Paper 82* .

Mascagni, Giulia, Christopher Nell and Nara Monkam. 2017. "One size does not fit all: a field experiment on the drivers of tax compliance and delivery methods in Rwanda." *ICTD Working Paper 58* .

McGraw, Kathleen M and John T Scholz. 1991. "Appeals to civic virtue versus attention to self-interest: Effects on tax compliance." *Law and society review* pp. 471–498.

Miguel, Edward and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1):159–217.

North, Douglas C. and Barry R. Weingast. 1989. "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England." *The Journal of Economic History* 49(04):803–832.

Ostrom, Elinor. 2000. "Crowding Out Citizenship." *Scandinavian Political Studies* 23(1):3–16.

Paler, Laura. 2013. "Keeping the Public Purse: An Experiment in Windfalls, Taxes, and the Incentives to Restrain Government." *American Political Science Review* 107(04):706–725.

Picur, Ronald D and Ahmed Riahi-Belkaoui. 2006. "The impact of bureaucracy, corruption and tax compliance." *Review of Accounting and Finance* .

Prichard, Wilson. 2015. *Taxation, responsiveness and accountability in Sub-Saharan Africa: the dynamics of tax bargaining.* Cambridge University Press.

Ross, Michael L. 2004. "Does Taxation Lead to Representation?" *British Journal of Political Science* 34(02):229–249.

Seim, Brigitte, Ryan Jablonski and Johan Ahlbäck. 2020. "How information about foreign aid affects public spending decisions: Evidence from a field experiment in Malawi." *Journal of Development Economics* 146:102522.

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

Stasavage, David. 2011. *States of credit: size, power, and the development of European polities.* Princeton University Press.

Tilly, Charles. 1992. *Coercion, capital and European states. AD 990-1992.* Cambridge MA and Oxford UK: Blackwell.

Timmons, Jeffrey F. 2005. "The Fiscal Contract: States, Taxes, and Public Services." *World Politics* 57(4):530–67.

Torgler, Benno. 2007. *Tax Compliance and Tax Morale: A Theoretical and Empirical Analysis.* Edward Elgar Publishing.

UNDP. N.d. Human Development Report. Technical report. Accessed June 2017.
**URL:** *http://hdr.undp.org/en/2016-report*

Weigel, Jonathan L. 2020. "The participation dividend of taxation: How citizens in Congo engage more with the state when it tries to tax them." *The Quarterly Journal of Economics* 135(4):1849–1903.

Weigel, Jonathan L and Elie Kabue Ngindu. 2023. "The taxman cometh: Pathways out of a low-capacity trap in the Democratic Republic of the Congo." *Economica* 90(360):1362–1396.

# Appendix Contents

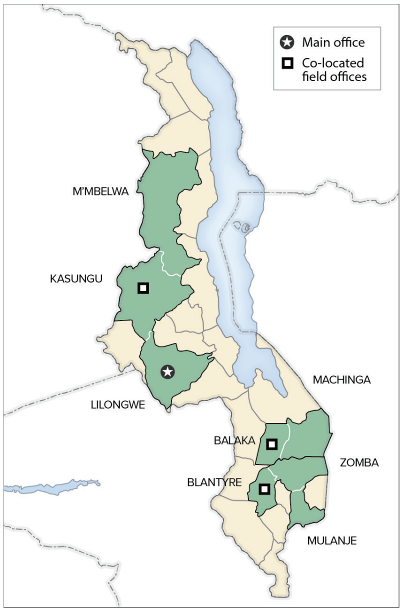


Figure A1: *Intervention Districts.* Main and Field offices refer to the offices of the USAID implementing partner.

# A Implemenation and Experimental Interventions

This section presents a detailed intervention timeline and supplemental details regarding each treatment component.

This field experiment was conducted as one component of the Local Government Accountability and Performance (LGAP) activity, a program from the United States Agency for International Development (USAID) in Malawi. USAID developed LGAP to support improved democratic accountability and local government capacity to effectively and efficiently deliver public services, for improved government performance. LGAP aimed to rigorously examine this link to support the Government of Malawi in determining the best ways to improve service delivery and democratic practice. It focused primarily on 1) Supporting citizen engagement and advocacy for accountable local government; 2) Building the capacities of local government to transparently deliver on their mandates; and 3) Supporting decentralization policy and process reforms as required by the Malawian 'Public Sector Reform' agenda. LGAP was implemeted by DAI over five years, August 2016-August 2021, with an initial total budget of approximately $15 million. It was implemented in 8 districts that were chosen by USAID and the Government of Malawi: Balaka, Blantyre, Kasungu, Lilongwe, Machinga, M'Mbelwa, Mulanje, and Zomba.

## A.1 Bottom-Up Treatment Bundle

**Step 1: Facilitate Market Committee Elections and Training**

At baseline, not all markets had valid market committees (or, indeed, any market committee at all). Invalid market committees are those that were formed without proper elections — such as committees that were directly installed in markets by the government — or whose terms have expired. Elections were held in the 54 treatment markets that did not have valid market committees. All newly elected market committees then received a training in which committee members learned about the proper organizational structure and the roles and responsibilities of the market committees, including their role in district governing structures. These trainings, co-run by each District Council's District Capacity Building staff and LGAP district staff, took place in December 2017 and January 2018.

**Step 2: Facilitate Meetings Between Vendors, Market Committees, and Local Government** After ensuring that each treatment market had a valid market committee, districts held a public meeting in the market to address vendors' sense of exclusion from the taxation system. These meetings included vendors, the market committee, political and bureaucratic district representatives, and group village headmen. District representatives typically in attendance were a representative from the District Finance Office, the market's tax collectors, the market/zone managers, and the ward councilor. A total of 3515 vendors attended these Step 2 meetings in the 64 treatment markets, with a median attendance of 61.5 vendors.[16] These meetings, which took place between January and February 2018, were observed by LGAP and included the following elements:

- A speech by the ward councilor in which they reminded vendors of the connection between taxes and development in the market
- A discussion of the roles and responsibilities of vendors and government officials, including vendors' obligation to pay market fees whether or not they sold any goods.
- A discussion of the perceived problems with the current tax collection system, in particular barriers faced by vendors. The district representatives also had the chance to explain how the revenue is used.
- An explanation of the bottom-up intervention and the way it will impact market operations. This included discussing the way the council uses funds from market fees and introducing Step 4 (the SMS system, see below). At the end of the meeting, vendors were able to register for the SMS system. A total of 2435 vendors signed up for the SMS system at the Step 2 meetings, with a median of 44 registrations. The median proportion of vendors who attended who signed up for the system was 0.729.[17]
- Documentation and discussion of the state of market services, including toilets, sanitation, security, and infrastructure. Vendors then selected their market development priorities. This was done using a pairwise ranking method over six options: market shed, borehole, electricity, pathways, a concrete slab, and refuse bins. This list was used to decide which infrastructure project a market would receive (see Step 3). This piece of the intervention in particular was meant to replicate tax bargaining.

**Step 3: Jump Start Service Delivery in Markets**

[16] These statistics exclude M'mbelwa markets; meeting reports are missing for this district.

[17] Excludes M'mbelwa markets, due to lack of data.

While our calculations suggested that market revenue would be sufficient to maintain infrastructure, existing compliance and funds were too low to fund infrastrcuture improvements. Step 3 therefore aided district governments to implement the market priorities chosen in Step 2. Forty-six markets chose a borehole – the other 18 chose a mix of the remaining project types. Projects cost approximately US$5,000, an amount insufficient to completely rehabilitate markets but sufficient to serve as a costly signal of the government's commitment to improving service provision in markets.

The markets were scoped and visited by construction specialists during the summer of 2018 in order to complete the necessary field assessments. A competitive bidding process between July and September 2018 led to the selection of the appropriate construction firms. Construction began in September 2018 and finished in March 2019. However, almost all markets saw at least some construction progress prior to endline data collection in November-December 2018. Each project was bookended by an opening ceremony and a handover ceremony, attended by government officials, market committees, and other vendors. Market committees were responsible for monitoring the state of the projects, and, upon their completion, developed a maintenance plan in conjunction with the district council.

**Step 4: Increase Transparency in Taxation via a SMS System**

To strengthen citizens' trust that their tax funds are being used well, and to facilitate transparency and bottom-up accountability between vendors and local government, Step 4 sought to provide citizens with information about government revenue and spending on an ongoing basis there an SMS messaging system. This system was developed and managed by mHub, an organization in Malawi that works with businesses and other organizations on information and communication technology projects(`http://www.mhubmw.com/`) At the meetings in Step 2, vendors were informed about the SMS system and were able to sign up by sending a SMS message to a market-specific number. Posters and pamphlets were also left at the markets after the meeting so vendors could understand the system better and sign up later.

Vendors then received monthly messages with that month's market revenue, along with information on how the money generally was allocated and spent. The text of these messages were designed to become more specific over the intervention period, as vendors became more comfortable with the system. One of the main advantages of the SMS system was that, once data on market revenues were collected, the system for passing this information to vendors was centrally managed and required few steps. Messages were first sent out in January 2018, in markets where Step 2 meetings had already taken place. The last messages were sent out in November 2018. All vendors who signed up received the SMS messages unless or until they opted out.

Vendors were also able to use a related SMS system, also set up and managed by mHub, to report grievances about local government service delivery. During the intervention period, grievances were passed on to district government officials designated by the implementing partner. mHub, in conjunction with the district governments, followed up with complainants when issues had been resolved. The grievance system was designed to give vendors more agency and enable them to take action if revenues were not used in line with their expectations or if market services were lacking.

## A.2 Top-Down Treatments

**Step 1: Roll Out Mobile Money Revenue Transfer System**
To address the widespread potential for evasion, corruption, and inefficiency at the market level, treatment markets shifted to transferring revenue via mobile money, a phone-based banking system. Airtel Malawi was engaged to transfer revenue to the district council. Tax collectors still collected fees from vendors and then gave the cash to the market manager. Then, instead of bringing the cash to the bank or district headquarters a few times a month, the market manager transferred the cash to an Airtel agent[18], who then transferred it to the district council bank accounts. Airtel earned 2% of the fees as payment. This streamlined revenue transfer to district governments and improved the government's ability to reliably track how much each market collects in fees. It also made it easier to see if certain markets were not transferring funds as regularly as they should. Markets started using the mobile money in March 2018 and continued using it until December 2018.

**Step 2: Provide Accurate and Reliable Market Vendor Counts**
One barrier to efficiently collecting market fees was the lack of a reliable estimate of anticipated revenue, which is required to determine collector benchmarks, monitor collector performance, and forecast local government revenue. However, the size of the market (measured in the number of vendors) changes over the course of the week, month, and year. This made a formal registration system unfeasible. Instead, the implementing partner hired and trained vendor counters.[19] Counters visited each market at least four times a month — on two market days and two non-market days. These vendor counters systematically walked through the market and recorded the number of vendors by type of business. On each visit, they counted vendors twice at different times of the day in order to obtain a more accurate count. Vendor counting started in February 2018 and continued until October 2018.

**Step 3: Forecast Revenue and Generate Revenue Targets Based on Vendor Numbers**
The figures produced in Step 2 were used to determine tax collector compensation schemes, forecast local government revenue, and track LGAP performance. The counts, once transferred to the government, were fed into a revenue target calculator that adjusted targets based on the previous month's revenues collected for each market and the number of market days a week that market had in order to create monthly estimates of the expected revenue for each market. These targets were then communicated to market masters and tax collectors. They were also used to evaluate market performance in Step 4. Targets were first sent to markets starting in April 2018. The last targets were communicated in November 2018.

**Step 4: Introduce Incentives for Tax Collectors**
Prior to the intervention, tax collectors reported low motivation due to low incomes, on average less than US$1 per day. This was true regardless of whether tax collectors received a fixed wage or commission pay.[20] This led to vendor perceptions of collector corruption.

[18] No market vendors were recruited as Airtel agents.

[19] These individuals could not be market vendors or government staff.

[20] Six districts reported at least some tax collectors paid based on commission for at least some of the

Salaries were also often late, reducing the incentive to work hard.[21]

Step 4 addressed these issues with an incentive system using the revenue targets created in Step 3. These incentives were non-monetary in nature and were applied at two levels: market and individual. If a market met or surpassed its monthly revenue target, the market team received either wheelbarrows, rakes, hoes, or shovels - valuable supplies that make management of the market easier. In addition, if the market met its target, each tax collector also received an individual incentive, which could have been a bicycle, fertilizer, certificates of excellence, mattresses, or work suites. A tax collector whose market kept meeting its targets was able to choose to alternate incentive goods. These incentives were designed to inspire tax collectors to perform their jobs without having to resort to bribery.

# B Deviations from Research Design and Intervention Plan

The treatments are described in their idealized form in Sections A.1 and A.2. Despite extensive buy-in from the implementing partners, district governments, and the national government, some markets saw significant deviations from the originally-planned intervention. This section catalogs the major issues in this section and address how these deviations may have affected our ability to detect results from the interventions. Because increasing taxation is almost always politically sensitive, the deviations we observe are also informative about the types of interventions that are most politically feasible in similar contexts.

## B.1 Project and Intervention Component Delays

The project as a whole was delayed several times. Baseline data collection began in July 2017. The goal at that point was to implement the interventions over the course of the next year, so that endline collection could take place in summer 2018 so that it would not overlap with the May 2019 elections. However, implementation delays meant that most of the top-down treatment components had only been in the field for 6-8 months during endline data collection. The fact that the markets and district government were only exposed to these components for approximately half of a year means that the effects we observe are likely the lower bound of potential top-down treatment impact.

In the the bottom-up treatment arm, the infrastructure component was significantly delayed for logistical reasons. Ultimately, only eight of the 64 projects started before October 30, when endline data collection began. Ultimately three markets were visited for endline data collection before any construction had taken place. Nineteen markets were visited after the handover ceremony had taken place and all construction had been completed.

implementation period.

[21] All districts except for Lilongwe experienced tax collector salary delays in at least some markets for some of the implementation period.

## B.2 Treatment Implementation Issues

In addition to general delays touched on in the previous section, there were a series of specific issues with treatment rollout that have the potential to weaken the interventions.

**Deviations in Bottom-up Markets:**
Due to budget constraints, about half of bottom-up markets ended up receiving their second-choice infrastructure project, not their first. The majority of these got their second choice. Additionally, despite scoping visits by hydrologists, no water was found after drilling in fourteen of the forty-six markets that were scheduled to receive a borehole. These markets received alternative projects, typically wheelbarrows and other cleaning supplies, or mobile refuse bins.

**Mobile Money Revenue System**:
In Balaka district, a dispute with the district's bank led to a temporary halt to the mobile money program for two months during the intervention period.

**Incentives for Meeting Targets (TD):**
The top-down treatment included individual incentives for tax collectors who met targets. However, after the first month of incentives, district governments switched the incentives to be market-level instead of individual. Following this change, markets that met their targets received market-level rewards of wheelbarrows and bicycles. Additionally, several districts did not consistently communicate revenue targets, limiting their impact. There was often always a significant lag before incentives were actually received by markets that met their targets.

## B.3 Protests, Boycotts, and Strikes

Throughout the course of the intervention period, a number of markets saw vendor protests and vendor boycotts of fee payments. These are common in the study context, and to the best of our knowledge were not the result of any intervention components.[22] Often, these protests had to do with lack of adequate services in markets, serving to underline the importance of the bottom-up treatment (particularly Step 3). Other protests were due to other market concerns unrelated to service provision.

# C Data Collection Strategy

The data in the main paper comes from baseline and endline surveys of market vendors and tax collectors. All surveys were implemented by Innovations for Poverty Action (IPA). Government revenue data, discussed below, was provided by our implementing partner, along with monthly reports on implementation. Finally, IPA performed monitoring visits to each

[22] There was one claim that the endline data collection team had set off protests in two markets; an independent investigation found that this was not the case, and the fee boycott was unrelated to the interventions.

of the sample markets, furnishing further data during the course of implementation. This all resulted in a rich tapestry of information upon which we draw in our analysis.

## C.1 Baseline and Endline Surveys

### C.1.1 Market Vendors Survey

Market vendors surveys attempted to sample 100 vendors from each of 128 markets; 20 for the full survey and 80 for the short version focusing on tax compliance. Different individuals were surveyed at baseline and endline, unless the same individual was chosen by chance. Vendors were selected using a modified random walk, details of which are available upon request. Markets were visited on their market days to ensure that the sample estimates reflected markets when the largest number of vendors were present. Vendors received a small airtime voucher in return for completing the survey, valued at USD0.28-USD0.42 for the short and long surveys.

### C.1.2 Tax Collectors Survey

Enumerators administered a 20-30 minute survey to up to seven tax collectors in each market. The survey included questions on knowledge of tax law; knowledge of customer service practices; number points of contact with market vendors and businesses; rejection rate in tax collection attempts; perceived proportion of market vendors paying taxes per day; amount collected in local taxes; and perceived barriers to tax compliance. At baseline, 302 tax collector surveys were completed, with an average of 2.44 per market. At endline, 264 tax collector surveys were completed, with an average of 2.06 per market.

## C.2 Monitoring Data

Some of our robustness checks investigate treatment noncompliance. These estimates rely on monitoring data from the intervention period between baseline and endline data collection. LGAP provided us with information on intervention status on a monthly basis. They also collected government records relating to tax collection on a monthly basis for the entire period between baseline and endline data collection. In addition, we carried out periodic focus groups with vendors and interviews with tax collectors, market committee members, and market managers to monitor implementation of the project.

### C.2.1 Data Exchange

The implementing partner, LGAP, provided key information that helped facilitate monitoring of the project's roll-out. In particular, this information helped us evaluate any spillover or violations to the project's planned interventions.[23] LGAP also served as the intermediary for data from the district governments, including market revenues. They also collected monthly market-level data on estimated numbers of vendors (in TD markets); numbers of tax collectors; the revenue targets (for TD markets); total revenue for each month; and LGAP activities in and around the market. More aggregated data were collected on how market revenues have been allocated/spent. In addition, LGAP provided district-level information on any treatment compliance issues, changes in tax collector and market manager employment, and any other potential spillovers or noncompliance issues.

### C.2.2 Market Visits

IPA also carried out unannounced market visits throughout the intervention period to supplement the quantitative analysis, allowing us to assess local perception of the interventions, provide an additional check of treatment compliance, and identify which mechanisms were being affected by the interventions. On average, 25 percent of the study markets were visited during every two-month period. Over the course of eight months, all 128 were scheduled to be visited. Each visited lasted about three hours, and included an anonymous walk in which the observer recorded observations, an interview with the market manager, an interview with the market committee chairperson, an interview with the tax collectors, and a one-hour focus group discussion with a small group of market vendors. These data provided us additional information on treatment compliance, spillovers, and how the interventions were perceived by different market actors.

## C.3 Measures

Table C1 presents the variables used to test hypotheses H1–H3, which correspond to the main hypotheses we laid out in Section 4.2. All measures are drawn from the data sources discussed in Section C.

# D Survey Descriptive Statistics

The tables in this section provide summary statistics for key demographics and outcomes in the baseline and endline surveys.

[23] In reality, this information was often delayed, which meant that we were often not able to react as quickly to issues as we would have liked. It was, however, invaluable to our analysis.

Table C1: Key Outcome Measures

| **Outcome** | **Question Wording** | **Answer Options** |
|---|---|---|
| receipt measure | Enumerators verified whether respondent could show a tax receipt and recorded the date on the receipt. Date used to confirm whether receipt dated from within past 7 days. | 1 / 0. |
| Self-reported compliance | Now, I am going to put 5 tokens on the table here. Think about the last 5 days you sold goods or services in this market. Please put a token here [indicate location] for each time you happened to be able to pay your K100 fee in the last 5 days | 0 - 5 |
| Perceived group compliance | Similar to self-reported compliance, but asked to allocate 10 tokens to represent other vendors paying fees. | 0 - 10 |
| Trust in Local Gov | In your opinion or based on what you have heard, would you say the district government is trustworthy? | 4-point scale |
| Trust in Ward Counc. | In your opinion or based on what you have heard,would you say your ward councilor for this market is trustworthy? | 4-point scale |
| Dist. Manages funds well | Do you strongly agree, agree, disagree or strongly disagree with how your district government is managing the following: Managing public funds effectively on behalf of citizens in this area | 4-point scale |
| District transparent spending | Do you strongly agree, agree, disagree or strongly disagree with how your district government is managing the following: Transparency in how it uses funds collected from market revenues | 4-point scale |
| District transparent collecting taxes | Do you strongly agree, agree, disagree or strongly disagree with how your district government is managing the following: Transparency in how much it is collecting from market revenues in this area | 4-point scale |
| Services Satisfaction: Combined | In general, how satisfied are you with the developments in THIS market provided by the district government? | 4-point scale |
| Satisfaction w Water | Now I am going to ask you about different developments in the market provided by the district government. For each, please tell me how satisfied you are with them. | 4-point scale |
| Perception of spending | For every 1,000 kwacha the government collects from this market, how much do you think goes towards developments in the market provided by the district government and paying market staff? | 0-1000 |
| Pay tax as a duty | Whether agree/disagree: “Paying taxes is a duty of all citizens, even when you do not approve of how elected officials spend money.” | 4-point scale |
| Tax morale | Which of these statements do you agree with more? Statement 1: Vendors should always pay tax even if they disagree with local government. Statement 2: Vendors should only pay tax if they agree with local government. | 4-point scale |

Table D1: Summary Stats for Demographic Variables Vendor Survey - Baseline. Superscripts in column names identify groups. Superscripts in cells indicate that a value is significantly different from the value for the superscripted group

| Variable | Overall Mean | BU$^{1}$ | BU & TD$^{2}$ | Control$^{3}$ | TD$^{4}$ | SD | Min | Max | N |
|---|---|---|---|---|---|---|---|---|---|
| age | 33.609 | 33.495 | 33.803 | 33.537 | 33.603 | 10.427 | 18.000 | 87.000 | 12356 |
| female | 0.334 | 0.355 | 0.342 | 0.302 | 0.337 | 0.472 | 0.000 | 1.000 | 12388 |
| educ_num | 8.268 | 8.387 | 8.117 | 8.23 | 8.336 | 3.436 | 0.000 | 15.000 | 12383 |
| literacy_any | 0.781 | 0.787 | $0.757^{4}$ | $0.761^{4}$ | $0.816^{2,3}$ | 0.414 | 0.000 | 1.000 | 2494 |
| hh_income_trim | 71512.791 | 73050.91 | 68903.4 | 72963.25 | 71156.43 | 86941.853 | 100.000 | 600000.000 | 11943 |
| service | 0.102 | 0.086 | 0.114 | 0.102 | 0.106 | 0.303 | 0.000 | 1.000 | 12388 |
| sell_daily | 0.283 | 0.297 | 0.282 | 0.269 | 0.283 | 0.450 | 0.000 | 1.000 | 12386 |
| yrs_in_mkt_fix | 6.402 | 6.199 | 6.5 | 6.354 | 6.561 | 6.383 | 0.000 | 47.000 | 2522 |

Table D2: Summary Stats for Demographic Variables Vendor Survey - Endline. Superscripts in column names identify groups. Superscripts in cells indicate that a value is significantly different from the value for the superscripted group

| Variable | Overall Mean | BU$^{1}$ | BU & TD$^{2}$ | Control$^{3}$ | TD$^{4}$ | SD | Min | Max | N |
|---|---|---|---|---|---|---|---|---|---|
| age | 33.974 | 33.828 | 34.165 | $34.334^{4}$ | $33.57^{3}$ | 10.140 | 18.000 | 86.000 | 12351 |
| female | 0.348 | 0.376 | 0.344 | 0.337 | 0.333 | 0.476 | 0.000 | 1.000 | 12370 |
| educ_num | 8.184 | 8.186 | 8.016 | 8.116 | 8.421 | 3.460 | 0.000 | 17.000 | 12358 |
| literacy_any | 0.845 | 0.864 | 0.854 | 0.838 | 0.825 | 0.362 | 0.000 | 1.000 | 2516 |
| hh_income_trim_99 | 62909.191 | 63105.49 | 60599.77 | 60989.77 | 66954.71 | 77482.333 | 1.000 | 600000.000 | 12159 |
| service | 0.087 | $0.059^{2,3,4}$ | $0.093^{1}$ | $0.094^{1}$ | $0.104^{1}$ | 0.282 | 0.000 | 1.000 | 12370 |
| sell_daily | 0.290 | 0.299 | 0.283 | 0.279 | 0.299 | 0.454 | 0.000 | 1.000 | 12369 |
| yrs_in_mkt_fix | 6.581 | 6.285 | 6.602 | 6.718 | 6.726 | 6.335 | 0.000 | 50.000 | 2525 |

Table D3: Summary Stats for Demographic Variables TC Survey - Baseline. Superscripts in column names identify groups. Superscripts in cells indicate that a value is significantly different from the value for the superscripted group

| Variable | Overall Mean | BU$^{1}$ | BU & TD$^{2}$ | Control$^{3}$ | TD$^{4}$ | SD | Min | Max | N |
|---|---|---|---|---|---|---|---|---|---|
| age | 41.523 | 41.932 | 41.905 | $43.838^{4}$ | $39.046^{3}$ | 11.895 | 20.000 | 88.000 | 302 |
| female | 0.265 | 0.274 | 0.23 | 0.221 | 0.322 | 0.442 | 0.000 | 1.000 | 302 |
| educ_num | 10.046 | $10.493^{3}$ | 9.878 | $9.338^{1,4}$ | $10.368^{3}$ | 2.453 | 0.000 | 14.000 | 302 |
| literacy_any | 0.973 | 0.973 | 0.959 | 0.971 | 0.988 | 0.161 | 0.000 | 1.000 | 301 |
| hh_income | 40263.907 | 40946.58 | 39777.03 | 39294.12 | 40863.22 | 37121.603 | 4000.000 | 350000.000 | 302 |
| days_wrk_mkt | 3.722 | $4.301^{2,4}$ | $3.554^{1}$ | 3.559 | $3.506^{1}$ | 2.251 | 1.000 | 7.000 | 302 |
| no_english | 0.358 | 0.397 | 0.338 | 0.294 | 0.391 | 0.480 | 0.000 | 1.000 | 302 |

Table D4: Summary Stats for Demographic Variables TC Survey - Endline. Superscripts in column names identify groups. Superscripts in cells indicate that a value is significantly different from the value for the superscripted group

| Variable | Overall Mean | BU$^{1}$ | BU & TD$^{2}$ | Control$^{3}$ | TD$^{4}$ | SD | Min | Max | N |
|---|---|---|---|---|---|---|---|---|---|
| age | 41.038 | $38.97^{3}$ | 42.322 | $43.141^{1}$ | 40.068 | 10.914 | 19.000 | 71.000 | 264 |
| female | 0.311 | 0.373 | 0.305 | 0.25 | 0.311 | 0.464 | 0.000 | 1.000 | 264 |
| educ_num | 10.163 | 10.06 | 10.254 | 10.016 | 10.311 | 2.232 | 3.000 | 13.000 | 264 |
| literacy_any | 0.924 | 0.91 | 0.915 | 0.906 | 0.959 | 0.265 | 0.000 | 1.000 | 264 |
| hh_income | 41328.939 | 38743.28 | 46891.86 | 40390.94 | 40045.95 | 31935.560 | 6000.000 | 250000.000 | 264 |
| days_wrk_mkt | 4.045 | $4.254^{3}$ | $4.475^{3}$ | $3.266^{1,2,4}$ | $4.189^{3}$ | 2.475 | 1.000 | 7.000 | 264 |
| no_english | 0.473 | 0.478 | 0.424 | 0.406 | 0.568 | 0.500 | 0.000 | 1.000 | 264 |

Table D5: Summary Stats for Outcome Variables Vendor Survey - Baseline. Superscripts in column names identify groups. Superscripts in cells indicate that a value is significantly different from the value for the superscripted group

| Variable | Overall Mean | BU$^1$ | BU & TD$^2$ | Control$^3$ | TD$^4$ | SD | Min | Max | N |
|---|---|---|---|---|---|---|---|---|---|
| fee1_full | 3.808 | 3.902 | 3.786 | 3.746 | 3.796 | 1.676 | 0.000 | 5.000 | 12359 |
| fee2_always | 6.689 | 6.811 | 6.7 | 6.57 | 6.672 | 2.506 | 0.000 | 10.000 | 12221 |
| recent_receipt_7 | 0.257 | 0.254 | 0.255 | 0.246 | 0.272 | 0.437 | 0.000 | 1.000 | 12372 |
| no_rcpt_when_pay_num | 1.357 | 1.344 | 1.337 | 1.367 | 1.383 | 0.783 | 1.000 | 5.000 | 2496 |
| tr1_num | 2.735 | 2.773 | 2.693 | 2.75 | 2.723 | 0.944 | 1.000 | 4.000 | 2463 |
| tr2_num | 2.674 | $2.794^{2,4}$ | $2.619^{1}$ | 2.669 | $2.609^{1}$ | 0.986 | 1.000 | 4.000 | 2413 |
| tr9e_num | 2.716 | 2.722 | 2.717 | 2.76 | 2.665 | 1.259 | 1.000 | 4.000 | 2495 |
| ms1_num | 2.000 | 2.05 | 1.963 | $1.861^{4}$ | $2.125^{3}$ | 1.241 | 1.000 | 4.000 | 2511 |
| ms3_num | 2.277 | 2.289 | 2.349 | 2.202 | 2.27 | 1.232 | 1.000 | 4.000 | 2506 |
| ms4_num | 2.486 | 2.502 | 2.57 | 2.418 | 2.456 | 1.185 | 1.000 | 4.000 | 2511 |
| ms5_num | 2.265 | 2.265 | 2.302 | 2.214 | 2.279 | 1.222 | 1.000 | 4.000 | 2509 |
| ms6_num | 2.597 | $2.694^{3}$ | 2.585 | $2.514^{1}$ | 2.589 | 1.272 | 1.000 | 4.000 | 2506 |
| satisfaction_dev_num | 2.046 | 2.111 | 2.082 | 1.961 | 2.03 | 1.143 | 1.000 | 4.000 | 12339 |
| ms_average | 2.328 | 2.363 | 2.359 | 2.243 | 2.347 | 0.867 | 1.000 | 4.000 | 2467 |
| tc2_10_clean | 301.970 | 308.407 | 303.375 | 284.094 | 311.771 | 216.008 | 0.000 | 1000.000 | 2361 |
| tax_morale_num | 1.544 | 1.541 | 1.556 | 1.536 | 1.542 | 0.498 | 1.000 | 2.000 | 12343 |
| tc2_4b_num | 3.691 | 3.654 | 3.727 | 3.689 | 3.695 | 0.694 | 1.000 | 4.000 | 2519 |
| tc5a_num | 1.483 | 1.45 | 1.476 | 1.537 | 1.471 | 1.006 | 1.000 | 4.000 | 2499 |
| tc5b_num | 1.749 | 1.722 | 1.71 | 1.792 | 1.771 | 1.151 | 1.000 | 4.000 | 2512 |
| tc2_15b_num | 3.825 | 3.818 | 3.834 | 3.839 | 3.807 | 0.506 | 1.000 | 4.000 | 2508 |
| petition | 0.708 | 0.709 | 0.707 | 0.724 | 0.692 | 0.455 | 0.000 | 1.000 | 2512 |
| petition_wname | 0.508 | 0.522 | 0.507 | 0.526 | 0.475 | 0.500 | 0.000 | 1.000 | 2518 |

Table D6: Summary Stats for Outcome Variables Vendor Survey - Endline. Superscripts in column names identify groups. Superscripts in cells indicate that a value is significantly different from the value for the superscripted group

| Variable | Overall Mean | BU$^1$ | BU & TD$^2$ | Control$^3$ | TD$^4$ | SD | Min | Max | N |
|---|---|---|---|---|---|---|---|---|---|
| fee1_full | 3.945 | 4.009 | 3.89 | $3.844^{4}$ | $4.034^{3}$ | 1.437 | 0.000 | 5.000 | 11822 |
| fee2_always | 6.508 | 6.661 | 6.508 | 6.392 | 6.473 | 2.357 | 0.000 | 10.000 | 12294 |
| recent_receipt_7 | 0.326 | $0.377^{3}$ | 0.325 | $0.265^{1}$ | 0.336 | 0.469 | 0.000 | 1.000 | 12365 |
| no_rcpt_when_pay_num | 1.477 | $1.417^{3,4}$ | $1.414^{3,4}$ | $1.533^{1,2}$ | $1.544^{1,2}$ | 0.828 | 1.000 | 5.000 | 2516 |
| tr1_num | 2.692 | 2.761 | 2.734 | 2.609 | 2.664 | 0.978 | 1.000 | 4.000 | 2509 |
| tr2_num | 2.600 | $2.715^{3,4}$ | $2.657^{4}$ | $2.555^{1}$ | $2.47^{1,2}$ | 1.006 | 1.000 | 4.000 | 2447 |
| tr9e_num | 2.516 | 2.457 | 2.477 | 2.545 | 2.587 | 1.128 | 1.000 | 4.000 | 2521 |
| tr9g_num | 2.382 | 2.316 | 2.363 | 2.409 | 2.441 | 1.151 | 1.000 | 4.000 | 2518 |
| tr9h_num | 2.362 | 2.307 | 2.323 | 2.401 | 2.419 | 1.162 | 1.000 | 4.000 | 2510 |
| ms1_num | 2.225 | $2.584^{2,3,4}$ | $2.239^{1}$ | $1.968^{1}$ | $2.103^{1}$ | 1.277 | 1.000 | 4.000 | 2517 |
| ms3_num | 2.356 | 2.461 | 2.282 | 2.359 | 2.323 | 1.230 | 1.000 | 4.000 | 2520 |
| ms4_num | 2.441 | 2.467 | 2.389 | 2.411 | 2.498 | 1.138 | 1.000 | 4.000 | 2520 |
| ms5_num | 2.188 | 2.144 | 2.141 | 2.142 | 2.325 | 1.153 | 1.000 | 4.000 | 2517 |
| ms6_num | 2.414 | 2.391 | 2.396 | 2.449 | 2.42 | 1.221 | 1.000 | 4.000 | 2525 |
| satisfaction_dev_num | 2.149 | $2.284^{3}$ | 2.17 | $1.999^{1}$ | 2.143 | 1.100 | 1.000 | 4.000 | 12365 |
| ms_average | 2.336 | 2.426 | 2.297 | 2.278 | 2.341 | 0.853 | 1.000 | 4.000 | 2478 |
| tc2_10_clean | 371.816 | 388.952 | 365.629 | 361.847 | 370.776 | 264.198 | 0.000 | 1000.000 | 2411 |
| tc9_clean | 724.017 | 717.399 | 737.372 | 722.463 | 718.709 | 260.398 | 0.000 | 1000.000 | 2463 |
| pay_even_disagree | 0.604 | 0.605 | 0.612 | 0.595 | 0.602 | 0.489 | 0.000 | 1.000 | 12355 |
| tc2_4b_num | 3.680 | $3.727^{3}$ | 3.68 | $3.638^{1}$ | 3.674 | 0.658 | 1.000 | 4.000 | 2531 |
| tc5a_num | 1.530 | 1.489 | 1.521 | 1.57 | 1.541 | 0.981 | 1.000 | 4.000 | 2514 |
| tc5b_num | 1.802 | 1.809 | 1.736 | 1.79 | 1.873 | 1.141 | 1.000 | 4.000 | 2524 |
| petition | 0.747 | $0.771^{4}$ | $0.781^{4}$ | 0.736 | $0.701^{1,2}$ | 0.435 | 0.000 | 1.000 | 2514 |
| petition_wname | 0.541 | $0.573^{4}$ | $0.583^{4}$ | 0.518 | $0.49^{1,2}$ | 0.498 | 0.000 | 1.000 | 2514 |
| stmt1_agree | 0.689 | 0.716 | 0.696 | 0.671 | 0.674 | 0.463 | 0.000 | 1.000 | 2531 |
| stmt1_agree_sent | 0.328 | $0.365^{4}$ | $0.359^{4}$ | 0.314 | $0.272^{1,2}$ | 0.470 | 0.000 | 1.000 | 2531 |
| stmt2_agree | 0.739 | 0.759 | 0.746 | 0.727 | 0.723 | 0.439 | 0.000 | 1.000 | 2531 |
| stmt2_agree_sent | 0.366 | 0.389 | 0.402 | 0.362 | 0.309 | 0.482 | 0.000 | 1.000 | 2531 |

Table D7: Summary Stats for Outcome Variables Tax Collector Survey - Baseline. Superscripts in column names identify groups. Superscripts in cells indicate that a value is significantly different from the value for the superscripted group

| Variable | Overall Mean | BU$^1$ | BU & TD$^2$ | Control$^3$ | TD$^4$ | SD | Min | Max | N |
|---|---|---|---|---|---|---|---|---|---|
| Hrs in Mkt | 9.405 | 9.71 | 9.617 | 8.996 | 9.289 | 2.595 | 1.000 | 14.000 | 302 |
| Vendors Visited | 93.282 | 100.443 | 78.716 | 67.552 | 119.724 | 187.634 | 10.000 | 3000.000 | 298 |

Table D8: Summary Stats for Outcome Variables Tax Collector Survey - Endline. Superscripts in column names identify groups. Superscripts in cells indicate that a value is significantly different from the value for the superscripted group

| Variable | Overall Mean | BU$^1$ | BU & TD$^2$ | Control$^3$ | TD$^4$ | SD | Min | Max | N |
|---|---|---|---|---|---|---|---|---|---|
| Hrs in Mkt | 9.841 | 9.276$^4$ | 10.359$^3$ | 9.218$^{2,4}$ | 10.495$^{1,3}$ | 3.002 | 0.500 | 20.250 | 262 |
| Vendors Visited | 113.923 | 77.621 | 178.966 | 72.922 | 130.347 | 434.360 | 12.000 | 6000.000 | 261 |

# E Revenue Results

Hypothesis 2 in our pre-analysis plan posited that each treatment would also increase the tax payments that reached district governments. We anticipated analyzing this outcome using market-level government revenue data, which was to be provided by the implementing partner on a monthly basis. We pre-specified that our analysis would include a difference-in-means estimate using November 2018 (endline) revenue data, as well as a diff-in-diff approach using November 2017 as the pre-treatment month.

However, there were significant issues with this data collection. Due in part to low capacity for record-keeping among district governments at baseline, we did not receive the November 2017 revenue numbers until May 2018, well after treatment was implemented. We cannot confirm that these numbers actually represented baseline revenues. For 17 disproportionately control markets, we never received any baseline revenue data at all. In contrast, later months were received much more promptly. This suggests that baseline data were especially prone to error, and that this may not be constant across treatment groups.

Figure E1 shows monthly market revenue estimates for each treatment group. For this analysis, the raw amount of revenue was divided by the daily fee payment, to create a more standardized measure of "market fee units per month." While there are noticeable differences at endline in the predicted direction, we cannot rule out that those differences were also present at baseline.

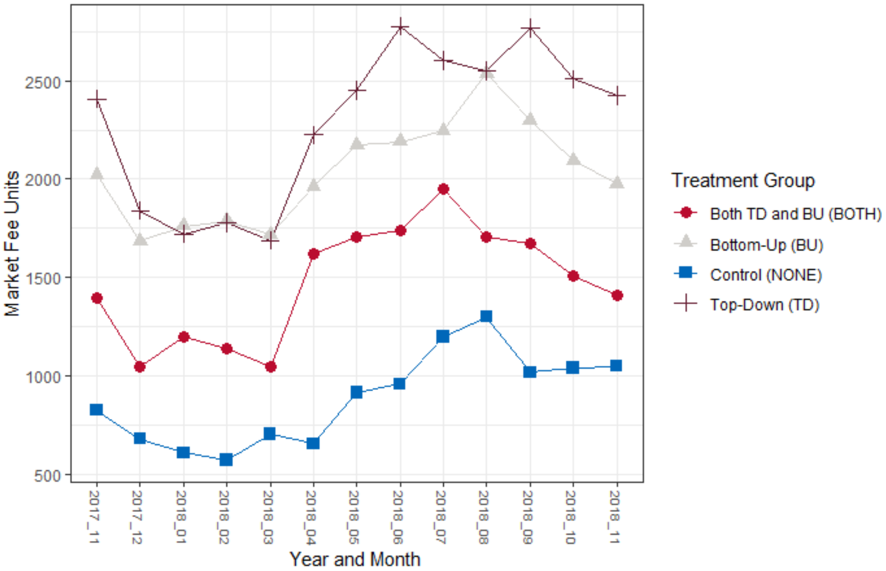


Figure E1: Market Revenue (Market Fee Units), Treatment Group Averages

Table E1 presents the DIM and DID estimates for the estimated number of monthly fee payments at the market level. The DIM regression shows a significant, positive effect of the TD treatment. However, as discussed above, the DID analysis, which uses November 2017 as a baseline, finds no significant treatment effects for any condition.[24]

[24] Using December 2017 as the baseline produces similar results.

Table E1: Hypothesis 2 Results Table, Market Fee Units

| | *Dependent variable:* | |
|---|---|---|
| | Market Revenue Collected<br>Market DIM | Market Revenue Collected<br>Market DID |
| BU | 961.004 | −94.445 |
| | (617.697) | (366.193) |
| TD | 1,251.142$^{*}$ | −154.090 |
| | (598.921) | (348.530) |
| Both | 341.216 | −159.338 |
| | (614.391) | (363.802) |
| Observations | 123 | 108 |
| Adjusted $R^2$ | 0.137 | −0.093 |
| Notes | | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001<br>Market-level models include block fixed-effects. |

These results have two possible interpretations. First, it is possible that the treatments really did increase government revenues, and that the baseline numbers are not accurately capturing pre-treatment revenue levels for the reasons described above. Second, it is possible that the treatment did not affect market revenues at all, and the DIM results reflect pre-existing differences in revenue. Note that the estimates also use slightly different samples, as we are missing baseline data for a number of markets. Ultimately, due to low confidence in the data, we have omitted these results from the main paper, focusing instead of measures where data quality is higher.

# F Understanding Differentiated Effects for BU & TD Treatment

In the results presented above, effects in the markets that received both the BU and the TD interventions (the BOTH group) are often either substantively smaller or statistically insignificant compared to the groups that received only one bundle of interventions (BU or TD). This pattern runs contrary to our expectation, which was that the two bundles would complement each other and, in combination, would have the greatest effect.

We posit four explanations for this pattern in the BOTH markets:

1. **Crowding Out Explanation:** In the BOTH markets, it is possible vendors were more inclined to pay taxes voluntarily due to the BU components, but this effect on voluntary tax compliance was counteracted ('crowded out') by the focus on consequences and

monitoring in the TD bundle. This explanation is supported in academic literatures from diverse fields (Agrawal, Chhatre and Gerber, 2015; Frey and Jegen, 2001; Ostrom, 2000).

2. **Vendor Capacity Explanation:** In the BOTH markets, it is possible that having eight intervention components roll out in one year was overwhelming for vendors, and their response was to ignore some of the components.

3. **State Capacity Explanation:** Planning, staffing, and managing all of the BU and TD components in the BOTH markets was resource-intensive, and it is possible district government delivered weaker versions of the treatments as a result.

4. **Intervention Timing Explanation:** The timing of the intervention rollout was such that some of the BU components (elections, meetings, SMS transparency campaign) rolled out before the TD interventions rolled out, and that many of the market infrastructure projects had not been completed prior to endline data collection. This means that, in the BOTH markets, vendors learned about their rights and responsibilities surrounding government revenue collection, then government focus on revenue collection was ramped up, but without the corresponding service improvements vendors were promised. This experience may have been particularly demoralizing for the vendors in the BOTH markets, especially in light of the Crowding Out Explanation (explanation (1)).

Without additional research, it is not possible to definitively determine which of these explanations is correct. Further, it is highly likely that all of these explanations are at play to some extent. Future work may need to focus on these issues in order to determine the most effective way to implement two-pronged interventions of this kind.

## G Mechanism Treatment Effects as Percent Increase over Control

Table G1 contextualizes the substantive significance of the treatment effects by presenting each as the percent increase over the control group mean.

Table G1: Treatment Effects as a Percent of Control Group

| **Outcome** | **Treatment Group** | **Treatment Effect** | **Percent Increase** | **Control Mean** |
|---|---|---|---|---|
| Trust in Local Gov. | BU | 0.176 | 6.745 | 2.609 |
| Trust in Local Gov. | Both | 0.142 | 5.432 | 2.609 |
| Trust in Ward Counc. | BU | 0.168 | 6.591 | 2.555 |
| Services Satisfaction | BU | 0.293 | 14.672 | 1.999 |
| Satisfaction with Water Access | BU | 0.654 | 33.220 | 1.968 |
| Satisfaction with Water Access | Both | 0.315 | 16.024 | 1.968 |
| Paying Tax as Duty | BU | 0.072 | 1.983 | 3.638 |
| Pay Because Consequences | TD | 0.056 | 1.526 | 3.642 |
| Money Flowing to Gov't | Both | 26.126 | 3.616 | 722.463 |
| Hours [TC] Working in Market A Day | TD | 1.154 | 12.521 | 9.218 |

# H Trust and Engagement: Additional Results

This appendix reports the additional analysis referenced in Section 6.3. Please note that Table H1 includes two unrelated models, combined in one table to save space.

The table shows that vendors in BU treatment markets who agreed to send a message against government overreach had a lower level of trust in the government than those who did not (a 7.55% decrease). In addition, the interaction between levels of trust in the local government (*tr1_clean*) and the BU only treatment group is statistically significant in the regression on whether vendors agreed with statement 1 in the whole sample, showing that individuals in this subgroup were actually slightly less likely to agree with statement 1, compared to vendors who had no trust at all in the local government.

Table H1: Political Engagement Outcomes

| | *Dependent variable:* | |
|---|---|---|
| | Trust in Local Gov | Agree St. 1 |
| stmt1_agree_sent | $-0.213^{***}$ (0.051) | |
| BU | | $0.122^{*}$ (0.057) |
| TD | | 0.014 (0.058) |
| Both | | 0.048 (0.061) |
| BU:tr1_cleanNot very trustworthy | | $-0.103$ (0.096) |
| BU:tr1_cleanSomewhat trustworthy | | $-0.023$ (0.065) |
| BU:tr1_cleanVery trustworthy | | $-0.153^{*}$ (0.072) |
| TD:tr1_cleanNot very trustworthy | | $-0.054$ (0.093) |
| TD:tr1_cleanSomewhat trustworthy | | 0.036 (0.069) |
| TD:tr1_cleanVery trustworthy | | $-0.103$ (0.073) |
| Both:tr1_cleanNot very trustworthy | | $-0.062$ (0.101) |
| Both:tr1_cleanSomewhat trustworthy | | 0.051 (0.068) |
| Both:tr1_cleanVery trustworthy | | $-0.018$ (0.072) |
| Observations | 1,263 | 2,509 |
| Adjusted $R^2$ | 0.011 | 0.318 |
| Notes: | | $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$<br>Models include enumerator and block fixed-effects.<br>Models have SEs clustered on market. |

# I Spillovers

## I.1 Introduction

Spillovers are a possibility in all experiments. In order to assess the extent to which treatment spillover is enhancing or diminishing the effects of the interventions, we employ two approachs: an inverse probability weighting (IPW) approach and a treatment externalities approach based on Miguel and Kremer (2004).

We use two approaches because the IPW approach, while canonical and useful, is somewhat of a poor fit for our situation because our treated units (markets) are a level higher than the observed units (vendors). Even when we use the individual level data on other markets in which vendors sell, we can only get an endline market level measure of spillover potential, not an individual level one, because we do not have panel data (see the next section for a more in-depth explanation). The treatment externalities approach allows us to take into account market size and how that may be impacting spillovers (termed treatment externalities by Miguel and Kremer (2004), hence the name).

## I.2 IPW Approach

With inverse probability weighting, "units are weighted by the inverse of the probability of being in the condition that they are in."[25] It requires making an assumption about where spillovers occur. In our case, we think about spillovers occurring geographically. If two markets are close to one another, it is possible the vendors from those markets actually visit or work in both markets. If the two markets have been assigned to different treatments, then those treatments may have "spilled over" between the two markets. For example, a vendor in a control market who sells in a bottom-up market may observe the infrastructure project there and may then have a similar reaction to a vendor in the bottom-up market.

We assume that spillovers will only occur within a certain distance around each market. We then use the distance between markets to create adjacency matrix. An adjacency matrix allows us to state mathematically whether individuals (or treated units) are connected (geographically, in our case) to another treated unit. We use the adjacency matrix to determine the actual treatment condition of a market—which is a mix of assigned and spillover conditions. There are 32 possible conditions, 8 each for each "pure" condition. For example, a market could be assigned to the bottom-up treatment, but they could be within $x$ km of another market that was assigned to the top down treatment. This market would then be in the "Bottom-Up_Top-Down" spillover condition group. We then simulate treatment assignment 10,000 times and calculate the number of times each market falls into each possible treatment condition. This gets us an estimate of the probability a market is in each possible treatment condition.

We use multiple adjacency matrices, which get us different probabilities and therefore different weights. A traditional adjacency matrix is an NxN indicator matrix, where N is the number of units, and where the cell [i,j] is 1 if unit i is adjacent to unit j and 0 otherwise.

We know the distance between each of our markets (except for Linjidzi, for which we are missing GPS coordinates). We also have information on where a portion (approx. 20%) of our sample sold in addition to the market in which they were interviewed.[26] We use three different versions of the adjacency matrix for IPW, combining these two data sources:

1. Distance only: a NxJ matrix, where N is the number of respondents and J is the number of markets – 1 if market j is within d distance of respondent's market, 0 otherwise.

2. Other Market Selling: a n X J matrix, where n is the number of respondents in our subsample (vendors who completed the long survey), and J is the number of markets. cell [i,j] is 1 if respondent i says they sell in market j.

3. Distance + Other Market Selling: this is once again an n x J matrix. We first add together the adjacency matrices for 1. and 2. If $A_{i,j_{1.}} + A_{i,j_{2.}} > 0$, cell [i,j] in this adjacency matrix takes on a value of 1. If 0, remains 0.

[25] https://egap.org/methods-guides/10-things-you-need-know-about-spillovers

[26] Vendors who completed the *long* survey were asked "Do you sell in any markets other than this one?" Those who responded *yes*, were then asked "what are the names of those markets?" An individual who noted a market within our sample was then "connected" to that market.

We use distances of two km, five km, and ten km. In each case, our results are only accurate if there is no spillover outside of that distance. In effect, this results in a sensitivity analysis: what happens when the spillover radius increases? With three distances and three different types of adjacency matrices, we end up with seven different adjacency matrices—2. above does not depend on distance.

We create 2. and 3. using baseline survey responses. We do this because we were concerned that responses to the question might have been affected by the intervention itself. To then incorporate this information into the endline analysis, we average the probabilities of being in the **modal** treatment condition among the market's respondents. This results in a market average. This means that for all models, all individuals within a market receive the same weights. We do this because we do not have a panel.

In our context, the IPW approach has some significant limitations. When we only use the distance between markets, we assume that *all* vendors are equally likely to go sell in nearby markets. Our data tell us this is very likely not the case. However, because we do not have a panel survey, when we incorporate individual responses, we are still forced to consider all individuals as having equal probability of being in the condition in which they are.

To account for spillovers in our main analysis, we first drop all markets that are currently in a spillover condition, and then weight individuals by the inverse of the probability that their market is in the pure treatment condition. We repeat this with the various probabilities calculated using our different adjacency matrices.

We do this for our main outcomes, with results shown in tables I1, I2, and I3.

Table I1: Spillover Analyses, Self-Reported Compliance

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Self-Rep. Comp. | | | | | | |
| | D2 | D2 - Mixed | D5 | D5 - Mixed | D10 | D10 - Mixed | Ind. Only |
| BU | 0.103 | 0.095 | 0.076 | 0.080 | 0.106 | 0.098 | 0.097 |
| | (0.079) | (0.093) | (0.079) | (0.093) | (0.115) | (0.113) | (0.092) |
| TD | $0.161^{*}$ | $0.193^{*}$ | $0.158^{*}$ | $0.194^{*}$ | $0.266^{*}$ | $0.314^{*}$ | $0.180^{*}$ |
| | (0.077) | (0.085) | (0.079) | (0.086) | (0.128) | (0.126) | (0.084) |
| Both | 0.021 | 0.064 | −0.038 | 0.028 | −0.030 | −0.002 | 0.063 |
| | (0.097) | (0.097) | (0.099) | (0.099) | (0.122) | (0.121) | (0.096) |
| Observations | 11,568 | 10,835 | 10,906 | 10,317 | 5,804 | 5,606 | 10,990 |
| Adjusted $R^2$ | 0.116 | 0.125 | 0.111 | 0.123 | 0.103 | 0.102 | 0.123 |

Notes: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

Table I2: Spillover Analyses, Group-Perceived Compliance

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Group-Per. Comp. | | | | | | |
| | D2 | D2 - Mixed | D5 | D5 - Mixed | D10 | D10 - Mixed | Ind. Only |
| BU | 0.154 | 0.049 | 0.127 | 0.024 | −0.139 | −0.043 | 0.048 |
| | (0.130) | (0.156) | (0.139) | (0.158) | (0.160) | (0.177) | (0.154) |
| TD | 0.024 | 0.023 | 0.021 | 0.018 | 0.100 | 0.145 | 0.011 |
| | (0.115) | (0.128) | (0.118) | (0.130) | (0.108) | (0.111) | (0.127) |
| Both | 0.023 | 0.051 | −0.038 | 0.019 | −0.255 | −0.189 | 0.047 |
| | (0.143) | (0.145) | (0.150) | (0.152) | (0.165) | (0.153) | (0.144) |
| Observations | 12,037 | 11,280 | 11,354 | 10,754 | 6,022 | 5,821 | 11,438 |
| Adjusted $R^2$ | 0.116 | 0.121 | 0.121 | 0.124 | 0.145 | 0.132 | 0.121 |

Notes: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

Table I3: Spillover Analyses, Evidence of Recent Receipt

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Evidence of Recent Receipt | | | | | | |
| | D2 | D2 - Mixed | D5 | D5 - Mixed | D10 | D10 - Mixed | Ind. Only |
| BU | 0.100$^{**}$ | 0.089$^{**}$ | 0.101$^{**}$ | 0.087$^{**}$ | 0.018 | −0.029 | 0.087$^{**}$ |
| | (0.032) | (0.032) | (0.033) | (0.033) | (0.047) | (0.043) | (0.032) |
| TD | 0.070$^{*}$ | 0.092$^{**}$ | 0.077$^{*}$ | 0.098$^{**}$ | −0.008 | 0.023 | 0.094$^{**}$ |
| | (0.031) | (0.031) | (0.032) | (0.031) | (0.032) | (0.023) | (0.031) |
| Both | 0.050 | 0.042 | 0.029 | 0.027 | −0.026 | −0.008 | 0.043 |
| | (0.032) | (0.033) | (0.031) | (0.032) | (0.037) | (0.039) | (0.032) |
| Observations | 12,108 | 11,348 | 11,422 | 10,820 | 6,037 | 5,836 | 11,506 |
| Adjusted $R^2$ | 0.264 | 0.260 | 0.288 | 0.279 | 0.317 | 0.285 | 0.265 |

Notes: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

## I.3 Treatment Externalities Approach

This approach is described in more depth in Miguel and Kremer (2004). We assume that spillovers are a function of the number of vendors or number of markets of a certain treatment condition within a certain distance from each market; the more vendors there are at nearby markets or, more roughly, the more markets there are a respondent's market, the more likely it is that the respondent will have heard about the treatment.

This amounts to fitting the following model:

$$\begin{aligned} Y_{ijkl} = \beta_0 + &\beta_1 * BU_j + \beta_2 * TD_j + \beta_3 * BOTH_j + \\ &\sum_d (\gamma_d * N_{dj}^{BU}) + \sum_d (\xi_d * N_{dj}^{TD}) + \sum_d (\zeta_d * N_{dj}^{BOTH}) + \sum_d (\phi_d * N_{dj}) + \\ &\beta_k * ENUM_k + \beta_l * Block_l + \epsilon_{ijkl} \end{aligned}$$

where $N_dj$ is the total number in markets at distance $d$ from market $j$, including market $j$ itself, and $N_{dj}^{BU}$, $N_{dj}^{TD}$, and $N_{dj}^{BOTH}$ are the numbers in markets assigned to the BU, TD, and BOTH treatments at distance $d$ from market $j$, respectively. To create the various $N_dj$, we add up

1. A daily average of the number of vendors who sell in a market
2. The maximum number of vendors who sell in a market during a week
3. The number of markets itself

We use the same distances as we do in the IPW approach: two km, five km, and ten km.

Table I4: Treatment Externalities, Self-Reported Tax Compliance

| | *Dependent variable:* | | |
|---|---|---|---|
| | | Self-Rep. Compliance | |
| | Avg. Vend. pr. Day | Max Num. Vendors | Num. Mkts. |
| BU | 0.118 (0.095) | 0.143 (0.097) | 0.204 (0.116) |
| TD | 0.118 (0.091) | 0.115 (0.093) | 0.059 (0.120) |
| Both | −0.046 (0.114) | −0.051 (0.115) | 0.014 (0.146) |
| N-2-BU | | | |
| N-5-BU | −0.002$^{***}$ (0.002) | −0.001 (0.001) | −0.184 (0.187) |
| N-10-BU | −0.00002 (0.0002) | −0.0001 (0.0001) | −0.114 (0.083) |
| N-2-TD | −0.002 (0.002) | −0.001 (0.001) | −0.359$^{***}$ (0.090) |
| N-5-TD | 0.005$^{***}$ (0.003) | 0.003 (0.001) | −0.022 (0.246) |
| N-10-TD | 0.0003 (0.0004) | 0.0001 (0.0002) | 0.131 (0.102) |
| N-2-Both | 0.001 (0.001) | 0.0005 (0.0004) | 0.210 (0.156) |
| N-5-Both | −0.0001 (0.001) | 0.00003 (0.0005) | 0.390$^{*}$ (0.186) |
| N-10-Both | 0.0003 (0.0002) | 0.0001 (0.0001) | −0.044 (0.098) |
| N-2-All | −0.001 (0.001) | −0.0004 (0.0004) | |
| N-5-All | 0.001$^{***}$ (0.001) | 0.001 (0.0004) | 0.138 (0.146) |
| N-10-All | −0.0003 (0.0001) | −0.0001$^{*}$ (0.00004) | −0.032 (0.068) |
| Observations | 11,623 | 11,623 | 11,623 |
| Adjusted $R^2$ | 0.117 | 0.118 | 0.120 |

Notes: $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

Table I5: Treatment Externalities, Group-Perceived Tax Compliance

| | *Dependent variable:* | | |
|---|---|---|---|
| | Group-Per. Compliance | | |
| | Avg. Vend. pr. Day | Max Num. Vendors | Num. Mkts. |
| BU | 0.094 (0.169) | 0.065 (0.168) | 0.068 (0.165) |
| TD | 0.052 (0.134) | 0.043 (0.142) | −0.002 (0.172) |
| Both | −0.036 (0.175) | −0.060 (0.178) | −0.111 (0.203) |
| N-2-BU | | | |
| N-5-BU | −0.002$^{***}$ (0.003) | −0.0003 (0.001) | −0.361 (0.304) |
| N-10-BU | 0.0003 (0.0002) | 0.0002 (0.0001) | 0.132 (0.128) |
| N-2-TD | 0.001 (0.004) | −0.001 (0.001) | −0.289$^{*}$ (0.146) |
| N-5-TD | 0.010$^{***}$ (0.005) | 0.005$^{*}$ (0.002) | 0.074 (0.396) |
| N-10-TD | −0.0002 (0.0005) | −0.00002 (0.0002) | 0.060 (0.157) |
| N-2-Both | 0.003 (0.002) | 0.001 (0.001) | 0.135 (0.285) |
| N-5-Both | −0.001$^{***}$ (0.002) | −0.0002 (0.001) | 0.336 (0.295) |
| N-10-Both | 0.0002 (0.0003) | 0.0001 (0.0001) | 0.112 (0.128) |
| N-2-All | −0.003 (0.002) | −0.001 (0.001) | |
| N-5-All | 0.003$^{***}$ (0.002) | 0.001 (0.001) | 0.354 (0.226) |
| N-10-All | −0.0001 (0.0001) | −0.00002 (0.00005) | −0.067 (0.089) |
| Observations | 12,096 | 12,096 | 12,096 |
| Adjusted $R^2$ | 0.117 | 0.117 | 0.117 |

Notes: $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

Table I6: Treatment Externalities, Evidence of Recent Receipt

| | *Dependent variable:* | | |
|---|---|---|---|
| | Evidence of Recent Receipt | | |
| | Avg. Vend. pr. Day | Max Num. Vendors | Num. Mkts. |
| BU | 0.059 (0.036) | 0.052 (0.036) | 0.077 (0.046) |
| TD | 0.021 (0.032) | 0.018 (0.033) | −0.002 (0.042) |
| Both | −0.001 (0.033) | −0.004 (0.034) | 0.002 (0.047) |
| N-2-BU | | | |
| N-5-BU | −0.001$^{***}$ (0.001) | −0.0005$^{*}$ (0.0002) | −0.235$^{***}$ (0.061) |
| N-10-BU | 0.0003 (0.0001) | 0.0001$^{***}$ (0.00003) | 0.044 (0.034) |
| N-2-TD | 0.002 (0.001) | 0.001$^{*}$ (0.0004) | 0.152$^{***}$ (0.040) |
| N-5-TD | −0.002$^{***}$ (0.001) | −0.001 (0.0005) | −0.365$^{***}$ (0.098) |
| N-10-TD | 0.0004 (0.0002) | 0.0002$^{*}$ (0.0001) | 0.106$^{*}$ (0.043) |
| N-2-Both | 0.001$^{*}$ (0.001) | 0.0005$^{**}$ (0.0002) | 0.286$^{***}$ (0.057) |
| N-5-Both | −0.001$^{***}$ (0.001) | −0.0002 (0.0002) | 0.073 (0.080) |
| N-10-Both | 0.0003 (0.0001) | 0.0001$^{**}$ (0.00003) | 0.037 (0.038) |
| N-2-All | −0.001$^{*}$ (0.001) | −0.0004$^{*}$ (0.0002) | |
| N-5-All | 0.001$^{***}$ (0.001) | 0.001$^{**}$ (0.0002) | 0.211$^{***}$ (0.050) |
| N-10-All | −0.0003 (0.00004) | −0.0001$^{***}$ (0.00001) | −0.029 (0.025) |
| Observations | 12,166 | 12,166 | 12,166 |
| Adjusted $R^2$ | 0.277 | 0.277 | 0.275 |

Notes: $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

# J Compliance Analysis

In this section, we estimate the so-called local average treatment effect (LATE), also known as the effect on compliers, using an instrumental variables strategy. We use treatment assignment as an instrument for treatment compliance. We operationalize treatment compliance in two ways, using the same set of compliance variables.

We say that a bottom-up treatment market has a compliance issue if it had one of three possible problems. First, if endline data collection occurred before mobilization for its infrastructure project had started. Second, if a vendor from a given market sent in a grievance message that was not responded to. Third, if a market did not receive the infrastructure project it had been promised just after mobilization began (there are multiple reasons for this, the most prevalent being that a borehole was drilled but no water was found).

We say that a top down treatment market has a compliance issue if it had one of three possible problems. First, if a market met its target but did not receive an incentive. Second, that the incentive a market was supposed to receive arrived delayed. Third, if mobile money was not active in any given month after May 2018.

We consider a market as having had a compliance issue under the *strict* operationalization when it had any one of the three issues. We consider a market as having had a compliance issue under the *relaxed* operationalization only when it had all three issues. The tables below present results for these models, focusing on our main outcome of the receipt measure.

Table J1: Compliance IV Regression 2nd-Stage Treatment Group Approach

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Self-Rep. Compl | | Group-Per. Compl. | | Recent Rcpt. | |
| | Strict | Relaxed | Strict | Relaxed | Strict | Relaxed |
| BU - Str. | 0.320 | | 0.566 | | 0.296$^{*}$ | |
| | (0.246) | | (0.400) | | (0.126) | |
| TD - Str. | 0.407 | | 0.131 | | 0.198 | |
| | (0.233) | | (0.332) | | (0.103) | |
| Both - Str. | −1.202 | | −0.897 | | −0.480 | |
| | (1.055) | | (1.512) | | (0.415) | |
| BU - Rel. | | 0.130 | | 0.214 | | 0.112$^{**}$ |
| | | (0.086) | | (0.144) | | (0.036) |
| TD - Rel. | | 0.179$^{*}$ | | 0.056 | | 0.083$^{*}$ |
| | | (0.085) | | (0.126) | | (0.034) |
| Both - Rel. | | 0.011 | | 0.052 | | 0.050 |
| | | (0.113) | | (0.173) | | (0.037) |
| Observations | 11,822 | 11,822 | 12,294 | 12,294 | 12,365 | 12,365 |
| Adjusted R$^2$ | 0.097 | 0.112 | 0.110 | 0.115 | 0.209 | 0.264 |

Notes: $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

Table J2: Compliance IV Regression 2-Stage Treatment Group Approach Market Level DIM

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Self-Rep. Compl | | Group-Per. Compl. | | Recent Rcpt. | |
| | Strict | Relaxed | Strict | Relaxed | Strict | Relaxed |
| BU Treat. - Str. | 0.233 | | 0.586 | | 0.285 | |
| | (0.354) | | (0.489) | | (0.152) | |
| TD Treat. - Str. | 0.456 | | 0.125 | | 0.147 | |
| | (0.324) | | (0.448) | | (0.140) | |
| BU Treat. - Rel. | −1.075 | | −0.507 | | −0.357 | |
| | (1.187) | | (1.640) | | (0.511) | |
| TD Treat. - Rel. | | 0.088 | | 0.222 | | $0.108^{*}$ |
| | | (0.123) | | (0.178) | | (0.048) |
| BU Treat. - Str. * TD Treat. - Str. | | 0.189 | | 0.052 | | 0.061 |
| | | (0.123) | | (0.178) | | (0.048) |
| BU Treat. - Rel. * TD Treat. - Rel. | | 0.012 | | 0.100 | | 0.052 |
| | | (0.129) | | (0.188) | | (0.051) |
| Observations | 128 | 128 | 128 | 128 | 128 | 128 |
| Adjusted $R^2$ | 0.087 | 0.235 | 0.058 | 0.132 | 0.331 | 0.535 |

Notes: $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001

Models include block fixed effects.

Table J3: Compliance IV Regression 2nd-Stage Treatment Group Approach Market Level DID

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Self-Rep. Compl | | Group-Per. Compl. | | Recent Rcpt. | |
| | Strict | Relaxed | Strict | Relaxed | Strict | Relaxed |
| BU - Str. | −0.220 | | −0.047 | | 0.291 | |
| | (0.489) | | (0.655) | | (0.173) | |
| TD - Str. | 0.303 | | −0.149 | | 0.090 | |
| | (0.449) | | (0.600) | | (0.159) | |
| Both - Str. | −0.334 | | −0.166 | | −0.328 | |
| | (1.642) | | (2.196) | | (0.582) | |
| BU - Rel. | | −0.084 | | −0.018 | | 0.110 |
| | | (0.170) | | (0.238) | | (0.056) |
| TD - Rel. | | 0.125 | | −0.062 | | 0.037 |
| | | (0.170) | | (0.238) | | (0.056) |
| Both - Rel. | | −0.034 | | −0.059 | | 0.046 |
| | | (0.179) | | (0.251) | | (0.059) |
| Observations | 256 | 256 | 256 | 256 | 256 | 256 |
| Adjusted $R^2$ | 0.080 | 0.227 | 0.163 | 0.233 | 0.429 | 0.582 |
| Notes: | | | | | | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001<br>Models include block fixed effects. |

# K Multilevel Modeling Analysis

Table K1: Multilevel Models

| | *Dependent variable:* | | |
|---|---|---|---|
| | Self-Rep. Compl | Group-Per. Compl. | Recent Rcpt. |
| BU | 0.073<br>(0.108) | 0.139<br>(0.166) | 0.091*<br>(0.038) |
| TD | 0.163<br>(0.108) | 0.050<br>(0.166) | 0.064<br>(0.038) |
| Both | 0.037<br>(0.108) | 0.057<br>(0.166) | 0.056<br>(0.038) |
| Observations | 11,822 | 12,294 | 12,365 |

Notes: $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001
Models include random intercepts by enumerators and by markets nested in blocks nested in districts.

# L Heterogeneous Treatment Effects Analysis

This section presents analysis of potential heterogenous treatment effects at the individual and market level. At the individual level we examine several vendor covariates.

## L.1 Market-level heterogeneity

At the market level we analyze market size (using baseline data) and a measure of vendors' collective action propensity. We measure collective action propensity by taking the market-level average of an endline survey question that asked “Do you agree or disagree with the following statement: When there is a problem in this market, we work together to solve it.”. Responses were on four-point scale that we then normalized; higher numbers indicate more agreement with the question.

## Table L1: Het. Treatment Effects by Market Size

| | | Self-Reported | | | Group-Perceived | | | Evidence of | |
|---|---|---|---|---|---|---|---|---|---|
| | | Compliance | | | Compliance | | | Recent Receipt | |
| | Mkt. DID | Mkt DIM | Ind. DIM | Mkt. DID | Mkt DIM | Ind. DIM | Mkt. DID | Mkt DIM | Ind. DIM |
| Market Size | 0.00001 | 0.00003 | 0.00004 | −0.00002 | −0.00002 | 0.000000 | 0.000004 | 0.00001 | 0.00001 |
| | (0.0001) | (0.0001) | (0.00004) | (0.0002) | (0.0001) | (0.0001) | (0.00004) | (0.00003) | (0.00003) |
| BU | −0.1167 | 0.0491 | 0.0992 | −0.1380 | 0.1331 | 0.1577 | 0.0863 | 0.0911* | 0.1071*** |
| | (0.1848) | (0.1337) | (0.0904) | (0.2679) | (0.1976) | (0.1717) | (0.0545) | (0.0510) | (0.0370) |
| TD | 0.1455 | 0.1434 | 0.1558* | −0.1095 | 0.0371 | 0.0761 | −0.0028 | 0.0022 | 0.0568 |
| | (0.1935) | (0.1400) | (0.0932) | (0.2805) | (0.2069) | (0.1399) | (0.0571) | (0.0534) | (0.0360) |
| Both | 0.0197 | 0.1431 | 0.1718* | −0.0436 | 0.1986 | 0.1905 | 0.0205 | 0.0323 | 0.0763** |
| | (0.1803) | (0.1304) | (0.1037) | (0.2613) | (0.1927) | (0.1629) | (0.0532) | (0.0497) | (0.0364) |
| Market Size * BU | 0.00005 | 0.00003 | 0.000004 | 0.0001 | 0.0001 | 0.00003 | 0.00001 | 0.00001 | −0.00001 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0002) | (0.0002) | (0.0001) | (0.00004) | (0.00004) | (0.00003) |
| Market Size * TD | −0.00004 | 0.00003 | 0.000003 | 0.0001 | 0.00002 | −0.00002 | 0.00005 | 0.0001 | 0.00002 |
| | (0.0002) | (0.0001) | (0.0001) | (0.0002) | (0.0002) | (0.0001) | (0.00005) | (0.00005) | (0.00003) |
| Market Size * Both | −0.0001 | −0.0002 | −0.0002*** | −0.00002 | −0.0002 | −0.0002** | 0.0001 | 0.00004 | −0.00003 |
| | (0.0002) | (0.0001) | (0.0001) | (0.0002) | (0.0002) | (0.0001) | (0.00005) | (0.00004) | (0.00003) |
| Constant | 0.4138 | 3.6563*** | 2.7310*** | 0.2484 | 6.4972*** | 6.1815*** | 0.0792 | 0.2379*** | 0.0139 |
| | (0.3258) | (0.2356) | (0.2380) | (0.4722) | (0.3482) | (0.2979) | (0.0961) | (0.0898) | (0.0392) |
| Observations | 128 | 128 | 11,822 | 128 | 128 | 12,294 | 128 | 128 | 12,365 |
| Adjusted $R^2$ | 0.0151 | 0.2515 | 0.1173 | −0.0053 | 0.1172 | 0.1166 | 0.2251 | 0.5704 | 0.2696 |

Notes: *p<0.1; **p<0.05; ***p<0.01
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.
Market-level models include block fixed-effects.

Table L2: Het. Treatment Effects by Collective Action Propensity

| | Self-Reported Compliance | Group-Perceived Compliance | Evidence of Recent Receipt |
|---|---|---|---|
| Collective Action Propensity | −0.046 | −0.024 | −0.037 |
| | (0.118) | (0.177) | (0.038) |
| BU | −0.186 | −0.151 | 0.097$^{*}$ |
| | (0.149) | (0.222) | (0.048) |
| TD | 0.076 | −0.104 | 0.044 |
| | (0.145) | (0.218) | (0.047) |
| Both | −0.031 | −0.120 | 0.046 |
| | (0.148) | (0.221) | (0.048) |
| Collective Action Propensity * BU | 0.465$^{**}$ | 0.504$^{*}$ | 0.085 |
| | (0.160) | (0.240) | (0.052) |
| Collective Action Propensity * TD | −0.193 | −0.236 | 0.032 |
| | (0.156) | (0.233) | (0.050) |
| Collective Action Propensity * Both | 0.145 | 0.348 | 0.103 |
| | (0.184) | (0.276) | (0.060) |
| Constant | 0.513 | 0.370 | 0.055 |
| | (0.312) | (0.466) | (0.101) |
| Observations | 128 | 128 | 128 |
| Adjusted R$^2$ | 0.177 | 0.105 | 0.218 |
| *Note:* | | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001 | |

## L.2 By Vendor Covariates

In this section, we fit only individual endline DiM models for the main outcomes, as vendor covariates are at the individual level. For binary variables (gender, service vs good stall type, selling daily) we fit subgroup models as well as interaction models.

Table L3: Subgroup Analysis by Gender

| | | Self-Reported | | | Group-Perceived | | | Evidence of | |
|---|---|---|---|---|---|---|---|---|---|
| | | Compliance | | | Compliance | | | Recent Receipt | |
| | Male | Female | Int. | Male | Female | Int. | Male | Female | Int. |
| Female | | | −0.093 | | | 0.060 | | | 0.088$^{***}$ |
| | | | (0.064) | | | (0.104) | | | (0.021) |
| BU | 0.144 | 0.105 | 0.137 | 0.140 | 0.303 | 0.143 | 0.097$^{**}$ | 0.090$^{**}$ | 0.100$^{**}$ |
| | (0.081) | (0.097) | (0.083) | (0.141) | (0.155) | (0.143) | (0.031) | (0.035) | (0.033) |
| TD | 0.150$^{*}$ | 0.182 | 0.145 | 0.036 | 0.092 | 0.015 | 0.093$^{**}$ | 0.036 | 0.091$^{**}$ |
| | (0.073) | (0.102) | (0.075) | (0.111) | (0.147) | (0.118) | (0.030) | (0.036) | (0.031) |
| Both | 0.035 | 0.047 | 0.031 | 0.029 | 0.111 | 0.021 | 0.058 | 0.041 | 0.061 |
| | (0.090) | (0.123) | (0.092) | (0.131) | (0.194) | (0.136) | (0.030) | (0.035) | (0.032) |
| Female * BU | | | −0.034 | | | 0.129 | | | −0.009 |
| | | | (0.090) | | | (0.139) | | | (0.030) |
| Female * TD | | | 0.038 | | | 0.110 | | | −0.048 |
| | | | (0.085) | | | (0.145) | | | (0.029) |
| Female * Both | | | 0.020 | | | 0.122 | | | −0.015 |
| | | | (0.089) | | | (0.157) | | | (0.030) |
| Constant | 2.834$^{***}$ | 2.655$^{***}$ | 2.823$^{***}$ | 6.377$^{***}$ | 5.863$^{***}$ | 6.243$^{***}$ | −0.067 | 0.201$^{*}$ | −0.008 |
| | (0.258) | (0.262) | (0.242) | (0.345) | (0.398) | (0.314) | (0.039) | (0.084) | (0.037) |
| Observations | 7,698 | 4,124 | 11,822 | 8,029 | 4,265 | 12,294 | 8,068 | 4,297 | 12,365 |
| Adjusted R$^2$ | 0.101 | 0.140 | 0.114 | 0.120 | 0.113 | 0.116 | 0.264 | 0.291 | 0.273 |
| *Note:* | | | | | | | | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001 | |

## Table L4: Subgroup Analysis by Services vs Goods

| | Self-Reported Compliance | | | Group-Perceived Compliance | | | Evidence of Recent Receipt | | |
|---|---|---|---|---|---|---|---|---|---|
| | Goods | Services | Int. | Goods | Services | Int. | Goods | Services | Int. |
| Service | | | −0.461*** | | | −0.148 | | | −0.040 |
| | | | (0.130) | | | (0.139) | | | (0.035) |
| BU | 0.088 | 0.465* | 0.086 | 0.161 | 0.688** | 0.160 | 0.101** | 0.010 | 0.101** |
| | (0.079) | (0.219) | (0.080) | (0.134) | (0.232) | (0.134) | (0.031) | (0.050) | (0.031) |
| TD | 0.163* | 0.128 | 0.156* | 0.062 | 0.059 | 0.059 | 0.077* | 0.011 | 0.078* |
| | (0.073) | (0.204) | (0.073) | (0.115) | (0.208) | (0.115) | (0.031) | (0.040) | (0.030) |
| Both | 0.033 | 0.042 | 0.029 | 0.075 | −0.262 | 0.076 | 0.059 | −0.009 | 0.060 |
| | (0.093) | (0.201) | (0.092) | (0.144) | (0.213) | (0.145) | (0.032) | (0.042) | (0.031) |
| Service * BU | | | 0.353 | | | 0.503* | | | −0.025 |
| | | | (0.189) | | | (0.218) | | | (0.048) |
| Service * TD | | | 0.066 | | | −0.070 | | | −0.033 |
| | | | (0.164) | | | (0.226) | | | (0.044) |
| Service * Both | | | 0.085 | | | −0.131 | | | −0.031 |
| | | | (0.177) | | | (0.201) | | | (0.047) |
| Constant | 2.966*** | 2.312*** | 2.874*** | 6.243*** | 6.118*** | 6.256*** | 0.043 | 0.081 | 0.032 |
| | (0.241) | (0.553) | (0.227) | (0.344) | (0.572) | (0.294) | (0.040) | (0.060) | (0.037) |
| Observations | 10,874 | 948 | 11,822 | 11,223 | 1,071 | 12,294 | 11,285 | 1,080 | 12,365 |
| Adjusted R$^2$ | 0.113 | 0.160 | 0.118 | 0.116 | 0.130 | 0.116 | 0.278 | 0.167 | 0.269 |

*Note:* $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001

## Table L5: Subgroup Analysis by Selling Daily or Not

| | Self-Reported Compliance | | | Group-Perceived Compliance | | | Evidence of Recent Receipt | | |
|---|---|---|---|---|---|---|---|---|---|
| | Not Daily | Daily | Int. | Not Daily | Daily | Int. | Not Daily | Daily | Int. |
| Sell Daily | | | −0.196 | | | 0.184 | | | 0.044 |
| | | | (0.113) | | | (0.108) | | | (0.035) |
| BU | 0.108 | 0.083 | 0.088 | 0.222 | −0.034 | 0.222 | 0.110*** | 0.040 | 0.113*** |
| | (0.075) | (0.150) | (0.080) | (0.127) | (0.199) | (0.133) | (0.033) | (0.040) | (0.034) |
| TD | 0.098 | 0.332* | 0.079 | 0.045 | 0.024 | 0.034 | 0.076* | 0.064 | 0.074* |
| | (0.067) | (0.145) | (0.071) | (0.115) | (0.170) | (0.122) | (0.033) | (0.039) | (0.033) |
| Both | 0.034 | 0.024 | 0.017 | 0.098 | −0.107 | 0.093 | 0.052 | 0.033 | 0.055 |
| | (0.079) | (0.189) | (0.082) | (0.132) | (0.223) | (0.132) | (0.032) | (0.040) | (0.032) |
| Sell Daily * BU | | | 0.126 | | | −0.113 | | | −0.046 |
| | | | (0.147) | | | (0.150) | | | (0.047) |
| Sell Daily * TD | | | 0.288* | | | 0.040 | | | −0.002 |
| | | | (0.138) | | | (0.166) | | | (0.046) |
| Sell Daily * Both | | | 0.078 | | | −0.110 | | | 0.002 |
| | | | (0.166) | | | (0.173) | | | (0.045) |
| Constant | 2.728*** | 2.891*** | 2.876*** | 5.556*** | 6.807*** | 6.140*** | −0.026 | 0.088 | −0.005 |
| | (0.355) | (0.291) | (0.243) | (0.375) | (0.325) | (0.294) | (0.068) | (0.052) | (0.040) |
| Observations | 8,491 | 3,330 | 11,821 | 8,719 | 3,574 | 12,293 | 8,780 | 3,584 | 12,364 |
| Adjusted R$^2$ | 0.119 | 0.150 | 0.114 | 0.128 | 0.098 | 0.116 | 0.288 | 0.249 | 0.269 |

*Note:* $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001

Table L6: Wealth Heterogeneous Effects Analysis

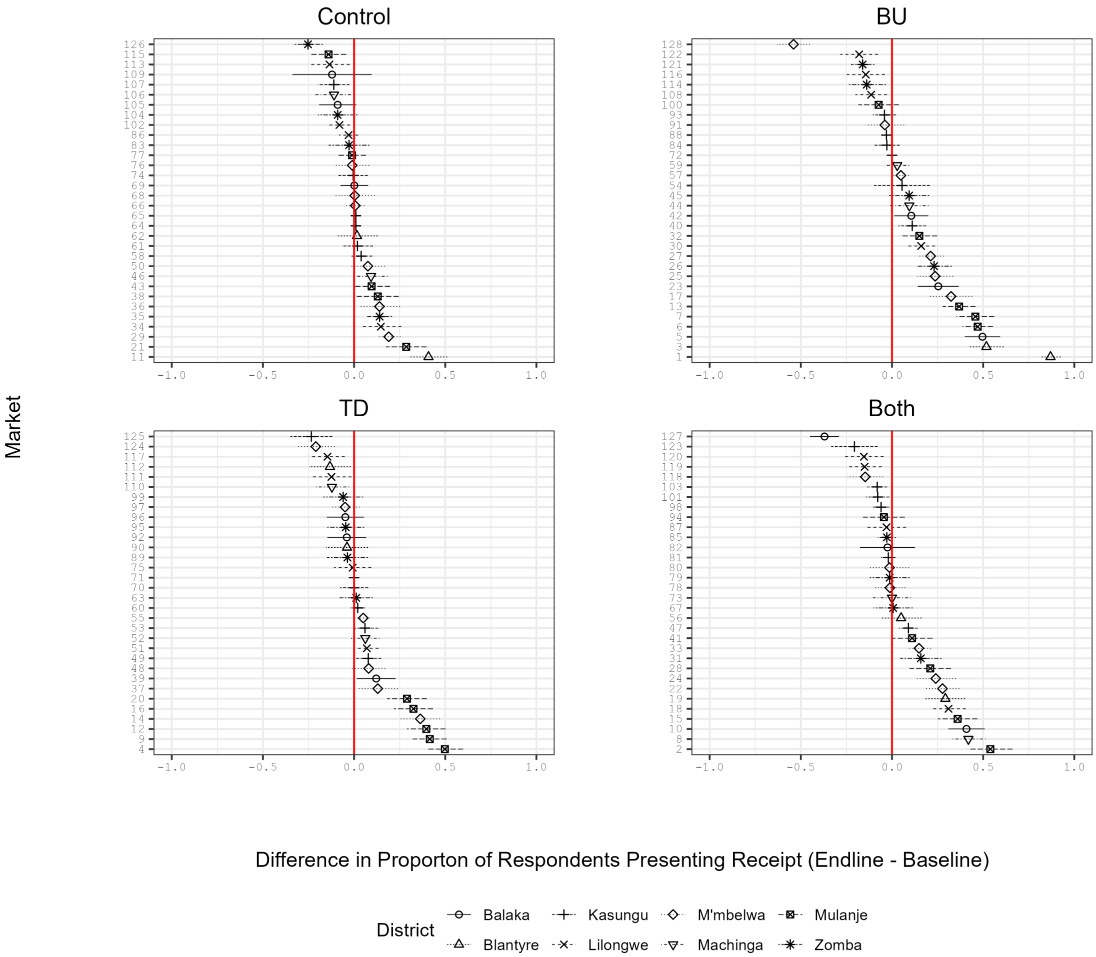
| | Self-Reported Compliance | Group-Perceived Compliance | Evidence of Recent Receipt |
|---|---|---|---|
| HH Income | 0.00000 | 0.00000 | −0.000 |
| | | (0.00000) | (0.00000) |
| BU | 0.080 | 0.193 | 0.108$^{***}$ |
| | (0.090) | (0.140) | (0.034) |
| TD | 0.138 | 0.077 | 0.082$^{***}$ |
| | (0.093) | (0.126) | (0.032) |
| Both | −0.019 | 0.095 | 0.071$^{**}$ |
| | (0.104) | (0.164) | (0.034) |
| HH Income * BU | 0.00000 | −0.00000 | −0.00000 |
| | | (0.00000) | (0.00000) |
| HH Income * TD | 0.00000 | −0.00000 | −0.00000 |
| | | (0.00000) | (0.00000) |
| HH Income * Both | 0.00000 | −0.00000 | −0.00000 |
| | | (0.00000) | (0.00000) |
| Constant | 2.766$^{***}$ | 6.206$^{***}$ | 0.015 |
| | (0.238) | (0.299) | (0.039) |
| Observations | 11,713 | 12,180 | 12,250 |
| Adjusted R$^2$ | 0.117 | 0.115 | 0.270 |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

# M General Robustness Models

## M.1 Main Outcomes

Figure M1: Evidence of Recent Receipt: Difference between Baseline and Endline by each market in each treatment group.



### M.1.1 Main Outcomes - Self-Report and Group-Perceived

Table M1: Hypothesis 1 Results Table - Individual-Level DIM and Market-Level DID

| **Panel A: Individual Level DIM Models** | | | |
|---|---|---|---|
| | Self-Reported<br>Full Tax Compliance | Perception of<br>Others' Always Complying | Evidence of Receipt<br>from Past 7 Days |
| BU | 0.119 | 0.194 | $0.101^{**}$ |
| | (0.079) | (0.132) | (0.031) |
| TD | $0.158^{*}$ | 0.050 | $0.074^{*}$ |
| | (0.075) | (0.114) | (0.030) |
| Both | 0.037 | 0.064 | 0.057 |
| | (0.094) | (0.142) | (0.031) |
| Observations | 11,822 | 12,294 | 12,365 |
| Adjusted $R^2$ | 0.113 | 0.115 | 0.268 |

| **Panel B: Market-Level DID Models** | | | |
|---|---|---|---|
| | Self-Reported<br>Full Tax Compliance | Perception of<br>Others' Always Complying | Evidence of Receipt<br>from Past 7 Days |
| BU | −0.076 | −0.016 | $0.100^{*}$ |
| | (0.150) | (0.217) | (0.045) |
| TD | 0.114 | −0.056 | 0.034 |
| | (0.150) | (0.217) | (0.045) |
| Both | −0.023 | −0.054 | 0.050 |
| | (0.150) | (0.217) | (0.045) |
| Observations | 128 | 128 | 128 |
| Adjusted $R^2$ | 0.049 | 0.024 | 0.211 |

Notes $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001
Individual-level models include enumerator and block fixed-effects
Individual-level models have SEs clustered on market.
Market-level models include block fixed-effects.

### M.1.2 Other Main Outcome Specifications

Table M2: H1: Self-Reported Tax Compliance Robustness Models

| | Self-Reported Full Compliance | | | |
|---|---|---|---|---|
| | Market DIM | Market DID | Individual DID | Individual DID Incl. Lagged DV |
| BU | 0.080 | 0.156 | 0.144 | 0.104 |
| | (0.111) | (0.108) | (0.088) | (0.076) |
| TD | 0.171 | 0.057 | 0.045 | 0.145** |
| | (0.111) | (0.108) | (0.103) | (0.073) |
| Both | 0.036 | 0.058 | 0.041 | 0.033 |
| | (0.111) | (0.108) | (0.080) | (0.093) |
| fee1_full_bl_avg | | | | 0.133 |
| | | | | (0.090) |
| Endline:BU | | −0.076 | −0.007 | |
| | | (0.153) | (0.118) | |
| Endline:TD | | 0.114 | 0.120 | |
| | | (0.153) | (0.124) | |
| Endline:Both | | −0.023 | −0.002 | |
| | | (0.153) | (0.133) | |
| Observations | 128 | 256 | 24,181 | 11,822 |
| Adjusted $R^2$ | 0.242 | 0.236 | 0.032 | 0.114 |

Notes: *p<0.1; **p<0.05; ***p<0.01
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.
Market-level models include block fixed-effects.

Table M3: H1: Group-Perceived Tax Compliance Robustness Models

| | Group-Perception of Always Complying | | | |
|---|---|---|---|---|
| | Market DIM | Market DID | Individual DID | Individual DID<br>Incl. Lagged DV |
| BU | 0.202 | 0.218 | $0.225^{**}$ | 0.207 |
| | (0.162) | (0.153) | (0.112) | (0.128) |
| TD | 0.047 | 0.103 | 0.091 | 0.060 |
| | (0.162) | (0.153) | (0.141) | (0.113) |
| Both | 0.104 | 0.158 | 0.127 | 0.071 |
| | (0.162) | (0.153) | (0.110) | (0.141) |
| fee2_always_bl_avg | | | | −0.055 |
| | | | | (0.083) |
| Endline:BU | | −0.016 | 0.013 | |
| | | (0.216) | (0.181) | |
| Endline:TD | | −0.056 | −0.047 | |
| | | (0.216) | (0.182) | |
| Endline:Both | | −0.054 | −0.020 | |
| | | (0.216) | (0.212) | |
| Observations | 128 | 256 | 24,515 | 12,294 |
| Adjusted $R^2$ | 0.129 | 0.231 | 0.025 | 0.115 |

Notes: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.
Market-level models include block fixed-effects.

Table M4: H1: Recent Receipt Robustness Models

| | Evidence of Receipt from Past 7 Days | | | |
|---|---|---|---|---|
| | Market DIM | Market DID | Individual DID | Individual DID Incl. Lagged DV |
| BU | 0.098** | −0.002 | −0.003 | 0.103*** |
| | (0.043) | (0.036) | (0.025) | (0.031) |
| TD | 0.055 | 0.021 | 0.021 | 0.064** |
| | (0.043) | (0.036) | (0.023) | (0.029) |
| Both | 0.056 | 0.007 | 0.005 | 0.053* |
| | (0.043) | (0.036) | (0.019) | (0.031) |
| recent_receipt_7_bl_avg | | | | 0.384*** |
| | | | | (0.149) |
| Endline:BU | | 0.100** | 0.102* | |
| | | (0.050) | (0.053) | |
| Endline:TD | | 0.034 | 0.037 | |
| | | (0.050) | (0.040) | |
| Endline:Both | | 0.050 | 0.049 | |
| | | (0.050) | (0.042) | |
| Observations | 128 | 256 | 24,737 | 12,365 |
| Adjusted $R^2$ | 0.551 | 0.594 | 0.153 | 0.271 |

Notes: *p<0.1; **p<0.05; ***p<0.01
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.
Market-level models include block fixed-effects.

### M.1.3 Interacting BU and TD Treatment Assignment

Table M5: Analysis as Factorial Design (w/ Int., no Int.)

| | Self-Reported<br>Full Compliance | | Self-Reported<br>Always Complying | | Evidence of Receipt<br>from Past 7 Days | |
|---|---|---|---|---|---|---|
| BU | 0.119 | $-0.002$ | 0.194 | 0.103 | $0.101^{***}$ | $0.041^{*}$ |
| | (0.079) | (0.061) | (0.132) | (0.094) | (0.031) | (0.022) |
| TD | $0.158^{**}$ | 0.038 | 0.050 | $-0.040$ | $0.074^{**}$ | 0.015 |
| | (0.075) | (0.060) | (0.114) | (0.094) | (0.030) | (0.022) |
| BU:TD | $-0.240^{**}$ | | $-0.180$ | | $-0.118^{***}$ | |
| | (0.119) | | (0.195) | | (0.043) | |
| Observations | 11,822 | 11,822 | 12,294 | 12,294 | 12,365 | 12,365 |
| Adjusted $R^2$ | 0.113 | 0.112 | 0.115 | 0.115 | 0.268 | 0.264 |

Notes: $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.

### M.1.4 0s as 0s for Self-Reported and Group-Perceived Tax Compliance

For the self-reported and group-perceived outcome measures, individuals were supposed to allocate 5 and 10 tokens, respectively, into three groups. The survey software was then supposed to check that all tokens had been allocated — enumerators should not have been able to proceed if allocations added up to less than 5 or 10. However, in some instances, the survey software seemingly malfunctioned, allowing respondents to report totals of more than or less than 5 or 10 for a single category or to report 0 for all categories. This was a larger problem for the self-reported compliance question, with ∼560 respondents dropping out of data. For group-perceived compliance, the number was smaller, at ∼100 respondents. In the main models, all of these were treated as NAs (that is, if an individual seemingly allocated none of their tokens or if they allocated more than 5 or 10 to a single category, that category's information was considered missing). To see what impact this may have had on our results we reran the data where only nonsensical (greater than 5 or 10 or less than 0) are treated as NAs, and *all 0* outcomes are retained.

Results are very similar. Tables omitted here to save space but are available on request.

### M.1.5 Alternative Outcomes

Table M6: H1: Evidence of Receipt from Past 10 Days

| | Evidence of Receipt from Past 10 Days | | |
|---|---|---|---|
| | Individual DIM | Market DIM | Market DID |
| BU | 0.105$^{***}$ | 0.106$^{**}$ | $-0.001$ |
| | (0.031) | (0.042) | (0.035) |
| TD | 0.076$^{**}$ | 0.054 | 0.024 |
| | (0.030) | (0.042) | (0.035) |
| Both | 0.057$^{*}$ | 0.055 | 0.007 |
| | (0.031) | (0.042) | (0.035) |
| Endline:BU | | | 0.107$^{**}$ |
| | | | (0.049) |
| Endline:TD | | | 0.030 |
| | | | (0.049) |
| Endline:Both | | | 0.048 |
| | | | (0.049) |
| Observations | 12,370 | 128 | 256 |
| Adjusted R$^2$ | 0.265 | 0.568 | 0.607 |

Notes: $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.
Market-level models include block fixed-effects.

Table M7: H1: Outcome 3 - Tax Collector Does **Not** Give You A Receipt When You Pay Fee

| | No Receipt When Paying | | |
|---|---|---|---|
| | EL DIM | BL-EL DID | EL DID<br>(Lagged DV) |
| Endline | | 0.052$^{***}$<br>(0.020) | |
| BU | $-0.059^{***}$<br>(0.020) | $-0.014$<br>(0.016) | $-0.059^{***}$<br>(0.020) |
| TD | $-0.004$<br>(0.019) | 0.010<br>(0.017) | $-0.003$<br>(0.019) |
| Both | $-0.054^{***}$<br>(0.019) | $-0.017$<br>(0.017) | $-0.055^{***}$<br>(0.019) |
| Endline:BU | | $-0.036$<br>(0.032) | |
| Endline:TD | | $-0.020$<br>(0.030) | |
| Endline:Both | | $-0.037$<br>(0.029) | |
| Observations | 2,516 | 5,012 | 2,516 |
| Adjusted R$^2$ | 0.105 | 0.009 | 0.105 |
| Notes | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01<br>All models have individuals as unit of analysis.<br>All include block fixed-effects.<br>Endline only models include enumerator fixed-effects as well.<br>All models have SEs clustered on market.<br>Lagged DV model includes baseline market average of DV.<br>Outcome is on binary. | | |

## M.2 Intermediate Outcomes

Table M8: Bottom-Up Causal Mechanism Outcomes: H4 - H5 - Individual-Level DID Results

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Trust Local Gov. | | Trust in Ward Cllr. | | DC Manages Funds Well | |
| | *OLS* | | *OLS* | | *OLS* | |
| | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) |
| Endline | −0.139$^{*}$ | | −0.109$^{*}$ | | −0.211$^{**}$ | |
| | (0.075) | | (0.060) | | (0.088) | |
| BU | 0.027 | 0.176$^{***}$ | 0.134$^{**}$ | 0.144$^{**}$ | −0.033 | −0.081 |
| | (0.058) | (0.063) | (0.066) | (0.069) | (0.067) | (0.058) |
| TD | −0.025 | −0.0004 | −0.055 | −0.108$^{*}$ | −0.091 | 0.004 |
| | (0.056) | (0.067) | (0.065) | (0.061) | (0.073) | (0.060) |
| Both | −0.057 | 0.137$^{**}$ | −0.043 | 0.109$^{*}$ | −0.039 | −0.054 |
| | (0.056) | (0.060) | (0.079) | (0.066) | (0.070) | (0.055) |
| tr1_bl_avg | | −0.077 | | | | |
| | | (0.104) | | | | |
| tr2_bl_avg | | | | 0.211$^{**}$ | | |
| | | | | (0.094) | | |
| tr9e_bl_avg | | | | | | 0.121 |
| | | | | | | (0.078) |
| Endline:BU | 0.124 | | 0.027 | | −0.055 | |
| | (0.102) | | (0.088) | | (0.114) | |
| Endline:TD | 0.074 | | −0.033 | | 0.134 | |
| | (0.112) | | (0.094) | | (0.126) | |
| Endline:Both | 0.184$^{*}$ | | 0.147 | | −0.029 | |
| | (0.103) | | (0.107) | | (0.113) | |
| Observations | 4,972 | 2,509 | 4,860 | 2,447 | 5,016 | 2,521 |
| Adjusted $R^2$ | 0.017 | 0.182 | 0.018 | 0.115 | 0.009 | 0.332 |

Notes: $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
All models include block fixed-effects.
Endline models include enumerator fixed-effects.
All models have SEs clustered on market.
Lagged DV models include market baseline average for DV.
All outcomes are on a 4-point scale.

Table M9: Bottom-Up Causal Mechanism Outcomes: H6 - Individual-Level DID Results

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Services Satisfaction | | Percep. of Sp. on Services | |
| | *OLS* | | *OLS* | |
| | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) |
| Endline | 0.037 | | $78.395^{***}$ | |
| | (0.047) | | (18.109) | |
| BU | 0.146 | $0.214^{***}$ | 23.733 | $24.396^{*}$ |
| | (0.093) | (0.071) | (16.199) | (14.631) |
| TD | 0.067 | 0.069 | $28.196^{*}$ | 3.165 |
| | (0.069) | (0.074) | (14.768) | (15.426) |
| Both | $0.121^{*}$ | 0.091 | 19.923 | $-2.189$ |
| | (0.073) | (0.075) | (14.956) | (13.931) |
| tc2_10_bl_avg | | | | 0.185 |
| | | | | (0.115) |
| Endline:BU | $0.133^{*}$ | | 2.329 | |
| | (0.078) | | (24.635) | |
| Endline:TD | 0.073 | | $-20.490$ | |
| | (0.084) | | (22.971) | |
| Endline:Both | 0.047 | | $-16.253$ | |
| | (0.075) | | (23.116) | |
| Observations | 24,704 | 12,365 | 4,772 | 2,411 |
| Adjusted $R^2$ | 0.033 | 0.196 | 0.034 | 0.291 |

Notes: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
All models include block fixed-effects.
Endline models include enumerator fixed-effects as well.
All models have SEs clustered on market.
Lagged DV models include market baseline average of DV.
Outcome 1 is on a 4-point scale. Outcome 2 is a number out of 1000.

Table M10: Bottom-Up Causal Mechanism Outcomes: H6 - Satisfaction with Specific Services

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Clean Water Access | Garbage Collection | Condition of Paths | Condition of Stalls | Security |
| | *OLS* | *OLS* | *OLS* | *OLS* | *OLS* |
| BU | 0.654*** | 0.149 | 0.073 | 0.022 | −0.035 |
| | (0.160) | (0.090) | (0.079) | (0.097) | (0.097) |
| TD | 0.161 | −0.041 | 0.073 | 0.150 | −0.060 |
| | (0.129) | (0.087) | (0.072) | (0.080) | (0.087) |
| Both | 0.315* | −0.045 | 0.004 | −0.010 | −0.034 |
| | (0.148) | (0.094) | (0.075) | (0.088) | (0.085) |
| Observations | 2,517 | 2,520 | 2,520 | 2,517 | 2,525 |
| Adjusted $R^2$ | 0.140 | 0.221 | 0.208 | 0.190 | 0.181 |

Notes: *p<0.05; **p<0.01; ***p<0.001
Individual-level models include enumerator and block fixed-effects.
Individual-level models have SEs clustered on market.
All outcomes are on a 4-point scale.

Table M11: Bottom-Up Causal Mechanism Outcomes: H6 - Satisfaction with Specific Services (Water Through Paths)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Clean Water Access | | Garbage Collection | | Condition of Paths | |
| | *OLS* | | *OLS* | | *OLS* | |
| | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) |
| BU | 0.133 | $0.575^{***}$ | 0.010 | 0.129 | 0.087 | 0.056 |
| | (0.126) | (0.155) | (0.095) | (0.083) | (0.088) | (0.074) |
| TD | $0.215^{*}$ | 0.045 | −0.012 | −0.058 | 0.043 | 0.068 |
| | (0.115) | (0.124) | (0.092) | (0.084) | (0.083) | (0.071) |
| Both | 0.058 | $0.263^{*}$ | 0.072 | −0.089 | $0.161^{**}$ | −0.034 |
| | (0.116) | (0.135) | (0.080) | (0.087) | (0.071) | (0.070) |
| ms1_bl_avg | | $0.429^{***}$ | | | | |
| | | (0.103) | | | | |
| ms3_bl_avg | | | | $0.309^{***}$ | | |
| | | | | (0.081) | | |
| ms4_bl_avg | | | | | | $0.271^{***}$ |
| | | | | | | (0.073) |
| BU:Endline | $0.538^{***}$ | | $0.168^{*}$ | | −0.034 | |
| | (0.154) | | (0.094) | | (0.099) | |
| TD:Endline | −0.014 | | 0.045 | | 0.032 | |
| | (0.101) | | (0.092) | | (0.095) | |
| Both:Endline | $0.270^{**}$ | | −0.070 | | $-0.186^{**}$ | |
| | (0.131) | | (0.095) | | (0.074) | |
| Observations | 5,028 | 2,517 | 5,026 | 2,520 | 5,031 | 2,520 |
| Adjusted $R^2$ | 0.052 | 0.168 | 0.042 | 0.229 | 0.036 | 0.214 |

Notes: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

All models include block fixed-effects.

Endline models include enumerator fixed-effects as well.

All models have SEs clustered on market.

Lagged DV models include market baseline average of DV.

All outcomes are on a 4-point scale.

Table M12: Bottom-Up Causal Mechanism Outcomes: H6 - Satisfaction with Specific Services (Stall Condition and Security)

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Condition of Stalls | | Security | |
| | *OLS* | | *OLS* | |
| | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) |
| BU | 0.081 | −0.0002 | 0.217$^{***}$ | −0.086 |
| | (0.086) | (0.084) | (0.067) | (0.090) |
| TD | 0.101 | 0.127$^{*}$ | 0.106 | −0.080 |
| | (0.081) | (0.076) | (0.079) | (0.085) |
| Both | 0.131$^{*}$ | −0.047 | 0.102 | −0.054 |
| | (0.076) | (0.080) | (0.081) | (0.080) |
| ms5_bl_avg | | 0.422$^{***}$ | | |
| | | (0.098) | | |
| ms6_bl_avg | | | | 0.288$^{***}$ |
| | | | | (0.106) |
| BU:Endline | −0.120 | | −0.306$^{***}$ | |
| | (0.076) | | (0.081) | |
| TD:Endline | 0.040 | | −0.176$^{**}$ | |
| | (0.097) | | (0.085) | |
| Both:Endline | −0.174$^{**}$ | | −0.186$^{**}$ | |
| | (0.074) | | (0.083) | |
| Observations | 5,026 | 2,517 | 5,031 | 2,525 |
| Adjusted $R^2$ | 0.026 | 0.204 | 0.015 | 0.186 |

Notes: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

All models include block fixed-effects and SE clustered by market. Endline models inc. enumerator FE.

Lagged DV models include mean market baseline DV. All outcomes on 4-pt scale.

Table M13: Bottom-Up Causal Mechanism Outcomes: H7 - Individual-Level DID Results

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Paying Tax as Duty | | Pay Tax Even if Disag. w. Gov. | |
| | *OLS* | | *OLS* | |
| | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) |
| Endline | −0.053 | | 0.060$^{***}$ | |
| | (0.054) | | (0.022) | |
| BU | −0.036 | 0.069$^{**}$ | 0.005 | 0.001 |
| | (0.047) | (0.034) | (0.016) | (0.012) |
| TD | 0.005 | 0.043 | 0.007 | 0.006 |
| | (0.045) | (0.030) | (0.017) | (0.011) |
| Both | 0.035 | 0.048 | 0.021 | 0.020 |
| | (0.046) | (0.033) | (0.016) | (0.013) |
| tc2_4b_bl_avg | | −0.151$^{**}$ | | |
| | | (0.063) | | |
| pay_even_disagree_bl_avg | | | | 0.130$^{*}$ |
| | | | | (0.074) |
| Endline:BU | 0.126$^{*}$ | | 0.005 | |
| | (0.072) | | (0.030) | |
| Endline:TD | 0.031 | | −0.001 | |
| | (0.069) | | (0.027) | |
| Endline:Both | 0.009 | | −0.003 | |
| | (0.067) | | (0.027) | |
| Observations | 5,050 | 2,531 | 24,698 | 12,355 |
| Adjusted R$^2$ | 0.004 | 0.112 | 0.006 | 0.082 |

Notes: $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

Individual-level models include enumerator and block fixed-effects. Individual-level models have SEs clustered on market.

Outcome 1 is on a 4-point scale. Outcome 2 is dichotomous.

Table M14: Top-Down Causal Mechanisms Outcomes, Vendor Survey - Individual-Level DID Results

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Could Refuse to Pay | | Group Non-Comp. Poss. | | Pay Because Consequences | |
| | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) | BL-EL DID | EL DID (Lagged DV) |
| BU | −0.091 | −0.043 | −0.071 | 0.026 | −0.020 | 0.035 |
| | (0.056) | (0.057) | (0.066) | (0.057) | (0.031) | (0.026) |
| TD | −0.068 | −0.046 | −0.022 | 0.065 | −0.033 | 0.052$^{**}$ |
| | (0.059) | (0.051) | (0.068) | (0.056) | (0.029) | (0.025) |
| Both | −0.065 | −0.039 | −0.085 | −0.045 | −0.006 | 0.041 |
| | (0.060) | (0.057) | (0.073) | (0.061) | (0.029) | (0.028) |
| tc5a_bl_avg | | 0.137 | | | | |
| | | (0.086) | | | | |
| tc5b_bl_avg | | | | 0.148$^{*}$ | | |
| | | | | (0.079) | | |
| tc2_15b_bl_avg | | | | | | −0.121 |
| | | | | | | (0.117) |
| Endline:BU | 0.010 | | 0.091 | | 0.048 | |
| | (0.091) | | (0.087) | | (0.064) | |
| Endline:TD | 0.040 | | 0.104 | | 0.069 | |
| | (0.085) | | (0.092) | | (0.053) | |
| Endline:Both | 0.018 | | 0.032 | | 0.033 | |
| | (0.093) | | (0.104) | | (0.055) | |
| Observations | 5,013 | 2,514 | 5,036 | 2,524 | 5,026 | 2,518 |
| Adjusted R$^2$ | 0.004 | 0.123 | 0.007 | 0.145 | 0.031 | 0.308 |

Notes: $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
All models include block fixed-effects.
Endline models include enumerator fixed-effects as well.
All models have SEs clustered on market.
Lagged DV models include market baseline average of DV.
All outcomes are on a 4-point scale.

Table M15: Top-Down Causal Mechanisms Outcomes, Tax Collector Survey – DID Results

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Hours Working in Market A Day | | Vendors Visited Per Day | |
| Endline | 0.271 | | −9.468 | |
| | (0.377) | | (15.163) | |
| BU | 0.607 | 0.089 | 34.680 | 35.928 |
| | (0.427) | (0.587) | (23.926) | (59.231) |
| TD | 0.319 | 0.921$^{*}$ | 44.153 | 103.900 |
| | (0.572) | (0.505) | (39.753) | (64.511) |
| Both | 0.537 | 0.309 | 18.215 | 189.129 |
| | (0.386) | (0.573) | (21.026) | (122.715) |
| hrs_in_mkt_bl_avg | | 0.297$^{***}$ | | |
| | | (0.098) | | |
| Endline:BU | −0.495 | | −11.485 | |
| | (0.633) | | (20.328) | |
| Endline:TD | 0.781 | | 13.731 | |
| | (0.841) | | (49.906) | |
| Endline:Both | 0.148 | | 103.736 | |
| | (0.629) | | (108.582) | |
| Observations | 566 | 260 | 559 | 257 |
| Adjusted $R^2$ | 0.235 | 0.379 | 0.089 | 0.308 |

Notes: $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

All models include block fixed-effects. Endline models include enumerator fixed-effects as well.
All models have SEs clustered on market. Lagged DV models include market baseline average of DV.
Outcome 1 is on a 4-point scale. Outcome 2 is dichotomous.

## N Multiple Hypothesis Testing Correction of Main Paper Results

As mentioned in our PAP, we performed multiple-hypothesis-correction to assess the robustness of our findings. Table N1 shows the *p*-values for the primary outcome models with significant treatment effects presented in the main paper. We present both the Holm correction, which controls the family-wise error rate (FWER), and the Benjamini-Hochberg correction, which controls the false discovery rate (FDR).

We combined all outcomes for each hypothesis. For the main hypothesis (H1), we also corrected for the fact that there are three comparisons against the control (evaluating the effect of the three treatment groups). For the BU mechanism and downstream (H4 - H8) and TD mechanism hypotheses (H9 - H11), we corrected only for the tests of the intervention effect relevant to each hypothesis (BU and Both, then TD and Both, respectively).

Table N1: Multiple Hypothesis Testing Correction

| Level | Outcome | Term | Hypothesis | $p$ | $p$<br>Holm | Survives<br>Holm |
|---|---|---|---|---|---|---|
| Individual | Self | TD | H1 | 0.034 | 0.504 | No |
| Individual | Receipt | BU | H1 | 0.001 | 0.025 | Yes |
| Individual | Receipt | TD | H1 | 0.014 | 0.241 | No |
| Market | Receipt | BU | H1 | 0.030 | 0.477 | No |
| Individual | Trust in Local Gov. | BU | H4 | 0.005 | 0.021 | Yes |
| Individual | Trust in Local Gov. | Both | H4 | 0.017 | 0.047 | Yes |
| Individual | Trust in Ward Counc. | BU | H4 | 0.016 | 0.047 | Yes |
| Individual | Services Satisfaction | BU | H6 | 0.002 | 0.012 | Yes |
| Individual | Paying Tax as Duty | BU | H7 | 0.038 | 0.152 | No |
| Individual | Petition Anon. | BU | H8 | 0.007 | 0.047 | Yes |
| Individual | Petition Anon. | Both | H8 | 0.004 | 0.042 | Yes |
| Individual | Petition w. Name | BU | H8 | 0.012 | 0.060 | No |
| Individual | Petition w. Name | Both | H8 | 0.007 | 0.047 | Yes |
| Individual | Agree St. 1 | BU | H8 | 0.006 | 0.047 | Yes |
| Individual | Agree St. 1 | Both | H8 | 0.004 | 0.042 | Yes |
| Individual | Agree St. 2 | Both | H8 | 0.022 | 0.088 | No |
| Individual | Pay Because Consequences | TD | H9 | 0.026 | 0.155 | No |
| Individual | Money Flowing to Gov't | Both | H10 | 0.020 | 0.040 | Yes |
| Individual | Hours Working in Market A Day | TD | H11 | 0.020 | 0.078 | No |

# O Explanation of Deviations from PAP

## O.1 Changes

A small change from the PAP is that we had specified that we would include district fixed effects. However, because we used block randomization, we used block fixed effects instead.

One of the major changes from the PAP is that we are treating the bottom-up and top-down treatment combination as its own treatment, for reasons already explained in the text.

We had specified that we would use a spillover radius of the largest distance a vendor seemed to travel, based on baseline responses. However, this turned out to be unreasonable: several vendors traveled more than thousand kilometers, according to our data (although this could be due to similar market names). The majority, however, sold only at one market, and so the mean distance traveled by a vendor was closer to zero. Therefore we settled on 2 km, 5km, and 10 km, before we did any spillover analysis.

# P Ethics

This appendix contains information on the ways in which this paper and the associated project comply with the ethical and transparency obligations described in APSA's A Guide to Professional Ethics in Political Science and in the APSA-approved Principles and Guidance for Human Subjects Research. Some of this information can also be found in prior appendices but is collected here for ease of access.

## P.1 Human Subjects Research

This project was evaluated and declared exempt from further review by the [UNIVERSITY REDACTED FOR ANONYMITY] IRB (IRB Number 17-1043) and by the NORC IRB (project number: 7554.030.01; IRB protocol number: 17.06.18). It was also evaluated and approved by the National Commission on Research in the Social Sciences and Humanities (which serves as the Malawi country-level IRB) as protocol P03/17/161.

We affirm that the study is in compliance with APSA's Principles and Guidance for Human Subjects Research. Please see below for more details.

**Consent and Confidentiality.** All interventions were conducted with the approval, and participation, of local government authorities. For the baseline and endline surveys, approval was given by local authorities, and by the market master. Enumerators asked each survey participant for their consent verbally before proceeding with the survey. During training, the necessity and importance of obtaining consent was repeatedly reinforced. The consent statement, which was approved by all relevant IRBs, stated that the survey was being conducted by Innovations for Poverty Action, and that the survey was about markets in Malawi. The voluntary nature of the study was stressed, as was confidentiality, and respondents were given contact information for IPA and the IRBs in case they had questions or concerns. We avoided mentioning the intervention in the consent process in order to avoid priming respondents about the intervention itself and to avoid additional social desirability bias when assessing tax compliance.

To promote confidentiality, enumerators were instructed to take respondents to a quiet place to complete the survey so that respondents' answers could not be overheard. Responses were stored on tablets during enumeration. At the end of each day, they were uploaded to IPA's server on the Box platform, where all data was encrypted using BoxCryptor. Only the PIs, their co-authors, and select IPA staff had access to the keys for this encrypted data. The data were also uploaded to a secure SFTP-protected server by NORC, to facilitate analysis. Only the PIs and their co-authors had access to this server. All publicly released data has been de-identified.

**Compensation.** At both baseline and endline, study participants who completed the vendor survey received a small airtime voucher in return for completing the survey (MWK200 for the short survey and either MWK300 or MWK600 for the long survey, depending on a delayed gratification experiment embedded in the long survey).

Participation in the surveys was not associated with any particular risks. Although several questions asked about tax compliance, which is technically required to operate in the market, the questions were phrased in such a way as not to ask directly about compliance or particular instances of compliance.

**Risk and ethical issues.** The interventions themselves were not assessed as causing undue risk by any of the IRBs that approved the project. It was technically possible that the top-down intervention would increase repression on behalf of the local authorities, although this would be an unintended side-effect as none of the top-down interventions were designed to give the authorities more power vis-a-vis market vendors. Checking for issues such as these were part of the ongoing monitoring assessments during the intervention period.

Our implementation partners did not report any ethical issues during the course of the study. There was one claim that the endline data collection team had set off protests in two markets; an independent investigation found that this was not the case, and the fee boycott was unrelated to the interventions.

**Funding.** This study was funded by the United States Agency for International Development. The implementing partner for the impact evaluation was NORC at the University of Chicago (initial project number: 7554.030.01; current project number: 7554.030). The implementing partner for the interventions in Malawi was DAI (Development Alternatives Incorporated). Surveys and monitoring were carried out by Innovations for Poverty Action Malawi.

**Data Collection Procedures, Data, and Code.** Please see Appendix C for information on data collection procedures. If the paper is accepted for publication, we will post all quantitative data and code necessary to replicate the results in the appropriate dataverse. The raw data will also be accessible through USAID's Development Data Library (https://data.usaid.gov/). The authors submitted the data to USAID through NORC at the University of Chicago on 4/20/22 but have no control over when USAID finishes processing.