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

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

Preliminary draft, please do not cite or circulate.

Abstract

Many developing countries appear stuck in a cycle of low state capacity: insufficient tax revenues limit public goods provision, but citizens will not pay taxes until governance improves. There is limited empirical evidence surrounding how to interrupt this cycle, especially whether it is more effective to improve the state's capacity to monitor and enforce taxation, or to instead strengthen taxpayers' motivation to pay taxes voluntarily. This paper presents the results of a field experiment on tax compliance and revenues conducted in 128 markets in Malawi. We use a 2x2 factorial design to compare the effectiveness of two capacity-building interventions: one that focused on improving top-down enforcement, and one that focused on improving accountability relationships and citizens' willingness to pay taxes voluntarily.

We find encouraging evidence that the interventions increased tax compliance. At endline, vendors in treatment markets were significantly more likely to have a tax payment receipt, which is the measure of tax compliance least vulnerable to measurement error. While we also find higher revenues in the top-down treatment markets at endline, we cannot eliminate the possibility that there were pre-treatment differences in revenue levels. Additional results suggest causal mechanisms. The bottom-up intervention bundle significantly increased vendors' satisfaction with services and their belief that paying taxes is a duty. In the top-down treatment group, vendors were more likely to report that they paid taxes because there were consequences if they did not, tax collectors report working longer hours, and vendors perceived lower bribe-taking. These findings suggest that local revenue collection can be kickstarted in a developing context, particularly by enhancing accountability and transparency surrounding tax revenue to, in turn, increase taxpayers' willingness to pay taxes.

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

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

[‡]Senior Research Scientist, International Programs, NORC at the University of Chicago

[§]Ph.D. Candidate, Department of Political Science, University of North Carolina at Chapel Hill

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 to best improve state capacity in a sustainable way. One aspect of state capacity that has been extremely stubborn to change is tax revenues. Even while GDP growth has increased across countries in sub-Saharan Africa and elsewhere, the percent of GDP taken in taxes has remained stable. Without more revenues, governments are unlikely to be able to expand the role of the state (Tilly, 1992; Stasavage, 2011; North and Weingast, 1989). This is also critical as taxation is thought to have a number of positive externalities, including promoting democracy, lowering corruption, and increasing citizen engagement; all of which could help sustain and expand state capacity further (Ross, 2004; Timmons, 2005; Baskaran and Bigsten, 2013; Brollo et al., 2013; Prichard, 2015; Paler, 2013; Weigel, 2017).

Expanding taxation is thus a critical element of improving state capacity. However, introducing new taxes will not produce revenue unless citizens actually comply. This makes it critical to understand how governments in developing countries can best increase the tax base by enforcing existing taxes. Existing theories suggest two key mechanisms through which this can occur. First, governments can invest in bureaucratic capacity and monitoring to decrease the costs of collection, including waste and corruption, and increase tax collector effort, thereby increasing both compliance and the percent of each dollar paid that goes to the government. Second, governments can bargain with citizens, providing public goods or other desired policies in return for at least quasi-voluntary compliance with taxation (Levi, 1989; North and Weingast, 1989; Bates and Lien, 1985; Prichard, 2015). In practice, most governments rely on a mixture of these two approaches.

We know little about which of these approaches works best in modern developing countries.

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. Theoretically, it is 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 government 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). It is also not clear whether enforcement and bargaining are complementary tactics, or whether using both is actually less effective than one approach alone.

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 tax compliance and government revenues 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" 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.

Markets are an excellent location to test compliance interventions. Market vendors are collected in dense, observable locations. This should make it feasible for governments both to collect revenues and provide local public goods. Market vendors also meet the preconditions for tax bargaining: there is broad agreement on how taxes should be spent, and vendors indicate high theoretical willingness to pay taxes, provided they see benefits in return, and vendors have high collective action potential.

Our experiment consists of two cross-cutting, market-level treatment bundles. The "bottom-up" intervention bundle was designed to improve quasi-voluntary compliance. It facilitated communication between market vendors and government; constructed new public goods in markets; and increased transparency regarding revenue levels and spending. The construc-

tion element makes this study one of the first to change *actual* levels of public services, rather than *perceptions* of these services. The"top-down" enforcement bundle aimed to improve local governments' ability to collect, track, and manage market revenue collection. It included improved revenue tracking technology (mobile money), improving government information about taxpayers, and incentive schemes for tax collectors.

We find that the bottom-up treatment significantly increases tax compliance among vendors, but has no effect on the amount of revenues that reach district government. In contrast, the top-down treatment has a less robust effect on compliance, but does significantly increase revenues. However, we cannot exclude the possibility that revenue differences also appeared at baseline. Critically, treatment effects are 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 interventions "crowded out" an increase in quasi-voluntary compliance from the bottom-up interventions, leading to a null effect on average.

Additional results provide evidence on intermediate mechanisms. The bottom-up interventions led to significant increases in vendors' trust in local government, satisfaction with services, and their belief that paying tax is a duty. In contrast, in the top-down treatment group, vendors were more likely to report that they paid taxes because there were consequences if they did not. We also find that tax collectors in the top-down markets report working significantly longer hours each day to collect taxes.

This paper makes several contributions to the literatures on state capacity and taxation. First, our results show that it is possible to increase state capacity through an intervention and, through it, increase tax compliance by jump-starting tax bargaining and a more positive taxation equilibrium. We also show that these effects are driven by the mechanisms proposed in the tax morale literature. However, we do not find strong evidence that these gains translate into higher government revenue. This suggests that, especially in the bottom-up

treatment group, there is still leakage of revenues to the government. While we do not have long-term outcomes, it is possible that any gains to compliance may be short-lived if they do not lead to higher government spending: increasing long-run tax compliance requires changing both citizen and government behavior.

Our work also advances on existing evidence along several fronts. In contrast to previous tax experiments, which typically target individuals, we treat entire markets, allowing us to test community-level determinants of taxation. 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, this research design allows us to test a key element of tax compliance theory that cannot be addressed through experiments that rely on treating individuals.

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 and government revenues in low-capacity states. It suggests, however, that weak or single-pronged approaches are unlikely to work; even our extremely strong interventions had limited success.

2 Theory

Each of our experimental treatments was designed to test a theoretical approach to tax compliance and state development. Our top-down treatment bundle is designed to increase compliance by improving local government's capacity to monitor taxation and tax collectors, and to improve tax collectors' incentives to work hard. This approach is based on theories that 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)). It also draws on standard principal-agent theory.

In practice, field experiments on taxation show that increasing the actual or perceived probability of detection and punishment can 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). Other recent studies suggest that incentivizing tax collectors can also improve tax compliance through higher tax collector effort (Khan, Khwaja and Olken, 2016; Weigel, 2018).

Our bottom-up treatment bundle, which is designed to improve market-government relations and trust, and to jump-start public goods provision, is based on theories 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).

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 to increase compliance, 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 are able 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).

Thus, citizens may pay taxes because they are compelled to, or because they feel that paying taxes is in line with their values and interests. In practice, most governments rely on a mix of the two approaches. Thus, our final treatment group gets both the top-down and bottom-up intervention bundles. Ex ante, it is unclear whether the two treatments will reinforce one another, or whether combining the two can have unintended negative effects on compliance. Relying on a mix of coercion and quasi-voluntary compliance could be more effective than either alone if, for example, using audits and penalties to compel those who don't pay voluntarily can bolster tax morale among those who do Coleman (2007). However, the opposite is also possible. Coercion could "crowd out" intrinsic motivation among those with high tax morale if public-minded citizens view coercion negatively and this lowers the perceived legitimacy of government (Dwenger et al., 2016).

Implementing both top-down and bottom-up approaches simultaneously could also backfire in low-capacity contexts like the one studied here if it leads to poor implementation and thus weaker treatments. Likewise, if vendors are overwhelmed by the number of changes to tax collection, they may simply ignore some intervention components. Finally, it may be the case that the *timing* of changes is important, in that improvements to trust and public services should precede any increase in perceive coercive pressure, to avoid crowding out effects. This theoretical ambiguity makes it critical to study the interventions together as well as separately.

2.1 Hypotheses

We expect each arm of our experiment to increase the fraction of vendors who pay market fees. We also expect each treatment to increase the revenue district governments receive from each market; this could be driven by higher compliance, or by more honest, efficient market employees.

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.

As discussed above, it is not obvious how the two treatment arms will interact. Our pre-analysis plan specified our prediction to be that:

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

Causal Mechanisms

A second set of hypotheses considers the causal mechanisms that increase tax compliance. If the bottom-up treatment works by increasing quasi-voluntary compliance and tax morale, we should observe that the treatment will:

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

The top-down treatment arm targeting coercive compliance is designed to make it easier for the government—and its representatives, including tax collectors, market managers, and district councilors—to collect taxes, while reducing corruption and shirking. We should therefore observe the treatment:

H8: Increase enforcement of the tax
H9: Decrease corruption
H10: Increase tax collector effort

3 Research Context

Malawi is representative of low-capacity states. Development is low, with an estimated 66.7% of the population multi-dimensionally poor (2014 UNDP Human Development Report). Over 37% of the government's budget comes from aid, and in 2013 Malawi ranked in the 34th percentile for government effectiveness (World Bank WGI). Local government capacity is especially weak. Local governments have significant authority over development but are almost entirely reliant on central government funding, which is generally insufficient to address local needs. From 2005 to 2014, districts were run entirely by bureaucratic appointees. Local elections were finally held in May 2014, but the newly-elected councilors are still working to establish systems and increase capacity.

Own-source revenues are critical for service provision but typically make up only a small percentage of rural districts' budgets. In many districts, 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 the local government in return is supposed to provide basic market services. Each day, tax collectors walk around the market to collect fees and give out receipts.[1] Collectors give revenues to the Market Master, who either deposit the cash or bring it to the district headquarters, typically about twice a month.

In our baseline sample self-reported tax compliance is 60%, but only 27% of respondents were able to produce a recent tax receipt, suggesting substantial over-reporting. In pre-treatment focus group interviews, vendors and market staff reported two barriers to higher tax compliance. First, vendors are unwilling to pay voluntarily. This is due to widespread dissatisfaction with market services, and a belief that tax revenues are spent on government salaries, rather than services. Vendors also reported that they, and the market associations

[1] Small markets may have a single part-time tax collector; large markets can have 20 or more full-time collectors.

that represent them, have been excluded from tax collection institutions and processes, and that this lowered tax compliance. Significantly, vendors consistently (if begrudgingly) acknowledged that the fee amount was *not* a barrier to compliance.

The second barrier to collection is low district capacity: some lack even a list of markets they collect fees from, and most had no data on market size. Because vendors pay in cash, this makes the fee collection process prone to corruption. A mix of low salaries and poor oversight means tax collectors have little motivation to work hard: at baseline tax collector incomes were \$0.80 to \$1.35 a day, low even in local terms.

4 Research Design

Our sampling frame consisted of 209 markets across the eight implementing partner target districts.[2] We 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. The appendix reports additional sample details.

Our field experiment randomized two cross-cutting treatment arms at the 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 efficiently. 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.

Because this is one of the first experiments to test the top-down and bottom-up mechanisms of improving tax compliance, we bundled several components together for each treatment. This ensures that any null effects are not due to weak treatment problems. It is also in line

[2]Balaka, Blantyre, Kasungu, Lilongwe, Machinga, Mulanje, M'mbelwa, and Zomba.

with our own qualitative fieldwork, which suggested that low compliance was due to multiple concurrent issues and thus any intervention needed to solve multiple "market failures" to increase compliance.

		Treatment 2: Top-Down Activities	
		Yes	No
Treatment 1: Bottom-up Activities	Yes	Group 1 32 markets	Group 2 32 Markets
	No	Group 3 32 Markets	Group 4 32 Markets

Table 1: 2x2 Factorial Experiment Design

4.1 Experimental Treatments

The components of each treatment were rolled out over a one-year period as as part of a larger, 5-year USAID program, LGAP, in the 8 sample districts. All concurrent LGAP components were designed to avoid confounding our analysis. 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 (October 2017- January 2018). To facilitate communication between markets and district government, market vendor committee elections were held in all markets without an active market association (54 markets total). These 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 (January - February 2018). Next, districts held public meetings in each market to address vendors' sense of exclusion from the tax system. In addition to vendors and their 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 funding to upgrade the market (step 3), and the transparency and grievance reporting system (Step 4). Vendors then chose their preferred market upgrades. 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 (September 2018 - March 2019). To escape the low services / low compliance equilibrium, all treatment markets received funding for a market infrastructure project, based on the priorities generated in Step 2. Projects cost approximately US$5,000. Each project was bookended by an opening ceremony and a handover ceremony, attended by government officials and vendors.

Step 4: Increase Transparency in the Taxation System via SMS Systems (February - December 2018). To facilitate ongoing communication and transparency between district governments and market vendors, we introduced an SMS system that vendors could sign up for during the Step 2 meetings. Each month, vendors received a message with information on the previous months' market revenue and how the money was allocated and spent. 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. Vendors received follow-up messages when an issue had been resolved. Seventy-three percent of kickoff meeting attendees signed up for the system.

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 reduce corruption on part of the tax collectors and market masters. The top-down treatment markets received four intervention components.

Step 1: Roll Out Mobile-Based Market Fee Payment System (April - December 2018). Tax collectors collect fees in cash from vendors and give the money to the market manager. In control markets, the manager remits cash to the district twice a month in person. As this system has high potential for corruption, in treatment markets, we introduced a mobile money system in which managers deposited tax payments daily into district bank accounts via a mobile money agent. This allowed districts to reliably track fee collection, and to detect gaps in payment.

Step 2: Provide Accurate and Reliable Market Vendor Counts (January - October 2018). 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.

Step 3: Generate Market Revenue Targets (May - November 2018).
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 revenue 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 (May - November 2018).
Finally, we implemented 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.[3][4]

5 Empirical Strategy

Our analysis uses data from baseline and endline surveys of vendors and tax collectors. These surveys were carried out by our data collection partner, Innovations for Poverty Action (IPA). In each market we surveyed 100 vendors at baseline and endline. Vendors were chosen via a modified random walk (see Appendix 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 causal mechanism and treatment compliance questions. 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.

We supplement the surveys with administrative data. Most critically, we have monthly tax collection information for each market in our sample, provided by the district government. We also draw on treatment compliance information from the implementing partner. Finally, we carried out periodic focus groups with vendors and interviews with tax collectors, market

[3]This component originally included individual incentives for each tax collector. These were eliminated after the first month due to district government concerns.

[4]In practice, incentives were frequently delivered late, and were not always given according to performance criteria. This may have weakened the top-down treatment.

committee members, and market managers to monitor implementation of the project.[5]

5.1 Outcomes Measurement

This section discusses the measures used to test our core hypotheses. For measures included only on the "long" vendor survey, we use individual-level analysis only. For measures that were also on the "short" survey, we also aggregate to form market-level estimates.

Hypothesis 1 predicts that each treatment should increase individual-level tax compliance. Our primary measure of tax compliance is the same as that used to do the block randomization: whether a vendor can produce a tax receipt from within the past 7 days. We use this as a verifiable measure that is less subject to response bias than self-reported measures. Our analysis uses both an individual- and market-level version of this variable.

For robustness we also draw on two additional measures. First, a self-reported measure of how many days, of the past 5 a vendor sold in that market, that they paid the entire fee. Second, a group-compliance measure in which respondents estimated the fraction of vendors (represented by 10 tokens) they believed always paid the fee.

To test Hypothesis 2, which predicted that each treatment would increase government revenues from each market, we use administrative data from district records. We divide reported monthly revenue for each market by the size of the market fee to create the number of fee payments per market in a given month. While we have this measure for every month between baseline and endline, we focus on November 2017 and 2018 as the baseline and endline months to allow for seasonal comparability.

Figure 1 plots the market fee units by treatment group throughout the intervention period, and shows substantial pre-treatment differences in average revenue. While we cannot completely rule out pre-treatment imbalance, Appendix XX shows that our sample is balanced

[5]See Appendix D for an in-depth explanation of data sources.

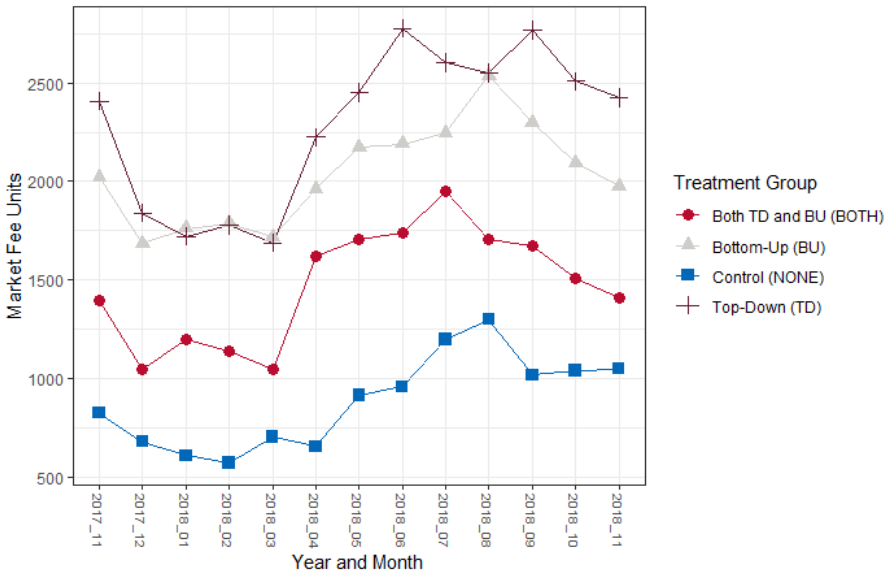

Figure 1: Market Revenue (Market Fee Units), Treatment Group Averages

on almost every attribute tested (the exception is a slight imbalance in tax collectors per market). Instead, these differences are likely due to two factors. First, the "baseline" month of November 2017 was after treatment assignment was revealed, and after some market committee elections had been held. Thus, we may not have a true pretreatment measure.

Second, we have reason to suspect significant measurement error at baseline in particular. All districts had much stronger revenue tracking and record-keeping by the end of the intervention period, due to other aspects of the larger LGAP project.[6] Indeed, districts were unable to produce the baseline (November 2017) revenue data until May 2018, well after treatments had rolled out; 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.

To test H3, which predicts that receiving both treatments will be superior to receiving either one, we use both the H1 and H2 outcome measures summarized above.

Our hypotheses section included a number of hypotheses about causal mechanisms for each

[6] As these improvements occurred for all markets regardless of treatment status, they are not a threat to inference.

treatment arm. To test these hypotheses, we use additional data on intermediate outcomes, drawn from the vendor and tax collector surveys. These outcomes are all measured at the individual, not market level, and are discussed in more detail below.

5.2 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 we initially planned (and put in our pre-analysis plan) to analyze 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. Appendix L.1.2 reports the interaction analysis.

Because our sample is not a panel (i.e., different individuals were interviewed at baseline and endline), 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.

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, β_1, β_2, and β_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 and ITT estimates represent the most conservative estimates.

6 Results

6.1 Treatment Effects: Tax Compliance

Hypothesis 1 predicted that each treatment would improve tax compliance. Table 2 reports the results for all three compliance measures described in Section 5.1. The top panel reports individual-level difference-in-means analysis, while the bottom reports the market-level difference-in-difference results.

The verified receipt measure—which was pre-registered as our primary measure—provides

the strongest evidence for an increase in tax compliance. 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. Those in the BOTH group were roughly six percentage points more likely to have a valid receipt, 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.

The self-reported compliance measure results (Column 1 of Table 2, Panel A) show that endline compliance is significantly higher in the TD treatment markets, but not the BU or BOTH groups. However, the these results do not hold once we account for baseline compliance levels in the difference-in-difference (DID) estimate (Panel B).In both the DIM and DID results we find no evidence that any treatment changed perceptions that other vendors are paying of the market fee (Column 2, Panels A and B). As the self-reported measures suffer from social desirability bias, especially the own compliance measure, we put more weight on our verified receipt measure.

To further explore the results, Figure 2 shows the market-level averages for the receipt measure. The four columns represent the four possible treatment assignments; each shape represents average market compliance at baseline (red) or endline (blue). The solid dots 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

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

Panel A: Individual Level DIM Models			
	Self-Reported Full Tax Compliance	Perception of Others' Always Complying	Evidence of Receipt 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 Full Tax Compliance	Perception of Others' Always Complying	Evidence of Receipt 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.

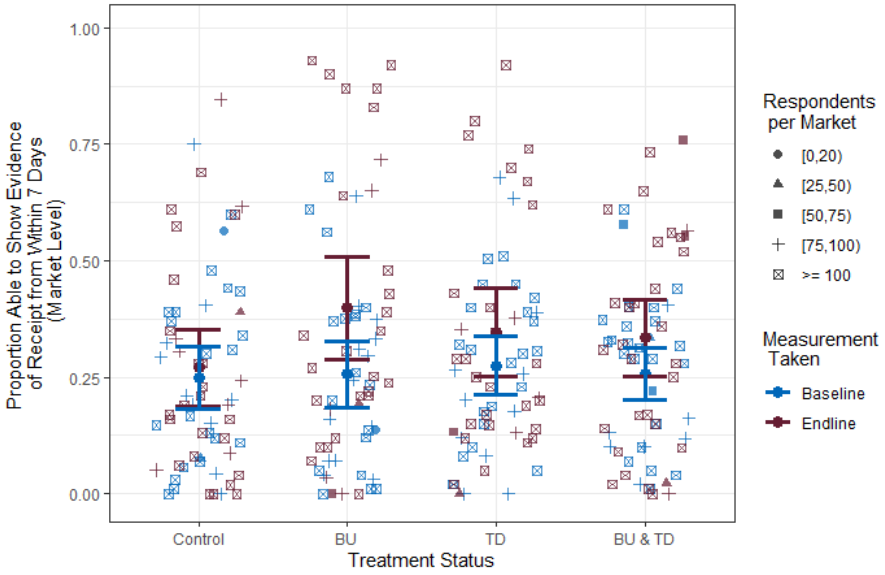

Figure 2: Evidence of Recent Receipt: Baseline and Endline

the regression results, Figure 2 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 groups, with the biggest change happening in the BU treatment group.

One concern is the receipt measure reflects vendors' ability to get a receipt, rather than higher compliance. To test this, Appendix Table 38 reports the results of a survey question in which we asked, on a 5-point scale, how often "you pay the fee but do not get a receipt". While we do find significant decreases in "pay but no receipt" in the BU and BOTH groups, the point estimate is too small (0.12 on the 5-point scale) to account for the ten 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, treatment effects are weakest in the markets that received both treatments; we discuss this pattern further below.

6.2 Treatment Effects: Revenue

Hypothesis 2 posited that each treatment would also increase the tax payments that reached district governments. Table 3 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, these differences existed in November 2017 as well. The DID analysis, which uses November 2017 as a baseline, finds no significant treatment effects for any condition.[7]

Table 3: Hypothesis 2 Results Table, Market Fee Units

	Dependent variable:	
	Market Revenue Collected Market DIM	Market Revenue Collected 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
		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 in Section 5.1. 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.

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

To further test whether revenue differences at endline reflect real treatment effects, we conducted time series analysis to examine the effects of the exact timing of different top-down treatment components. Because qualitative data suggested that the mobile money component was the strongest, we ran by-month within-market analysis that uses market-level variation in when mobile money started and stopped in each treatment market. We do not detect higher revenues in market-months in which mobile money was active.

6.3 Causal Mechanism Effects

Our theory suggested a number of causal mechanisms that could affect the tax compliance and revenue outcomes. We posited that the BU treatments would increase tax compliance through their effects on trust (H4), satisfaction with government and with services (H5 and H6), and tax morale (H7). To test these hypotheses, we use additional questions from the endline survey (see Appendix for details), analyzed in Tables 4 and 6. As predicted, we find that the BU treatment increased trust in the district government and ward councilor, and that treated vendors were more likely to view paying taxes as a duty, one of our operationalizations of tax morale.

Finally, Table 6 also shows that vendors exposed to the BU treatment were more satisfied with market services in general than those in the control group. This result is driven by a large increase in satisfaction with access to clean water, likely because boreholes were the chosen construction project in 43 of 64 BU treatment markets.

We interpret these results as supporting the intended causal mechanism for the BU treatment: in interacting more with district government officials and experiencing more responsiveness and transparency 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 4: Bottom-Up Causal Mechanism Outcomes: H4 - H5

	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

Table 5: 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.

In the TD intervention, we predicted that the treatment bundle would increase perceived and actual tax enforcement (H8); decrease corruption (H9); and increase tax collector effort (H10). Tables 7 and 8 report results from the vendor and tax collector surveys, respectively (see Appendix for question wording).

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, 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.

	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				

Table 6: 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.

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.

To further measure the treatments' effects on corruption, we included a list experiment on 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. If we group all intervention groups together, the list experiment estimates that 14% of tax collectors report accepting money from a vendor seeking to avoid paying the market fee. If we estimate corruption for each treatment group, 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 these estimates suggest that the TD treatment in particular may have reduced rent-seeking, we lack statistical power to test whether these differences are statistically significant.

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 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.

	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

Table 7: 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.

Table 8: 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 just described show that both the top-down and bottom-up treatments affected state capacity. The bottom-up treatment increased tax compliance, but these additional revenues do not appear to have reached the government. The top-down treatment had a smaller impact on tax compliance, but may have increased government revenues at least somewhat. In each case, the results suggest that the effects are driven by the expected theoretical mechanisms. 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.

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 endling difference-in-

means, market-level difference-in-differences, individual-level quasi-difference-in-differences[8], and individual-level endline difference-in-means controlling for the baseline market-level average for the outcome variable (see App. **??**). We also analyzed the main outcomes using an interaction between top down and bottom-up treatment arms (in other words, analyzing the 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. L.1.3), and widening the receipt window to ten days (see Table 47 in App. L.1.4). As an alternative to the receipt measure, we also tested the interventions' impact on whether a tax collector does not give a receipt; the bottom-up and both treatment groups show a negative effect here, backing up out results (see Table 48 in App. L.1.4).)

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. L.2). We also dichotomized key causal mechanism outcomes (ones that were significant in the original analysis), which corroborate the causal mechanism analyses presented here (see App. L.2.3).

The revenue analysis shows the same patterns in some different specifications and variable formulations, including comparing November 2017 to November 2018 in the difference-in-differences, and treating markets for which we did not have revenue information in a certain month as having 0 revenue that month (see App. H.2). However, the treatment group differences disappear if we look at market fees and market fee units per tax collector and do not survive a difference in difference analysis. Logging the outcome variable does not change these conclusions.

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

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

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 Section J. 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.

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-government relations, jump-starting tax bargaining, and providing a way out of the low-services, low-compliance tax equilibrium. The second top-down arm focused on enforcement and government capacity to monitor and collect taxes efficiently.

Both interventions succeeded in increasing tax compliance, with the largest effects in the bottom-up treatment arm. The top-down treatment may have increased revenues, but poor baseline data make this result less clear. However, we see no real evidence of treatment effects in markets that received both treatment arms. This suggests that, especially in low-capacity states, trying to implement too many forms 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.

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* .

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.".

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.".

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.

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.

Weigel, Jonathan L. 2017. "Building state and citizen: How tax collection in Congo engenders citizen engagement with the state." *Harvard University, Cambridge, MA* .

Weigel, Jonathan L. 2018. "The taxman cometh: A virtuous cycle of compliance and state legitimacy in the DR Congo." *unpublished typescript* .

Appendix Contents

A USAID LGAP Project

This field experiment was conducted as one component of a larger USAID project, LGAP. USAID developed the Local Government Accountability and Performance (LGAP) activity to support improved democratic accountability and local government capacity to effectively and efficiently deliver public services, for improved government performance. The aim of LGAP is to rigorously examine this link to support the Government of Malawi in determining the best ways to improve service delivery and democratic practice. LGAP focuses primarily upon three areas: 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 Public Sector Reform (PSR) agenda.

LGAP is being rolled out over five years, will involve many implementing partners (including our primary point of contact for the tax intervention), and will cost approximately $15 million. Because LGAP is a large, multi-year project with multiple components, there were several design issues that had implications for this project. First, all LGAP activities were happening throughout the implementation period of the field experiment. We were concerned these activities, particularly the ones being discussed that may provide information to citizens about government performance or focus on revenue accountability in non-market areas, may "wash out" the effect of the randomized tax compliance treatments across treatment and control markets.[9] We were also concerned that the heightened activity in local government would inhibit our ability to isolate the control markets.

LGAP was conducted in 8 districts in Malawi, shown in Figure 3. These districts were chosen by the implementing partner; all districts are either USAID target districts, or districts chosen by the Government of Malawi (GOM) as the pilot districts for the next round of

[9]LGAP activities began in October 2016, approximately a year before the interventions associated with the project started.

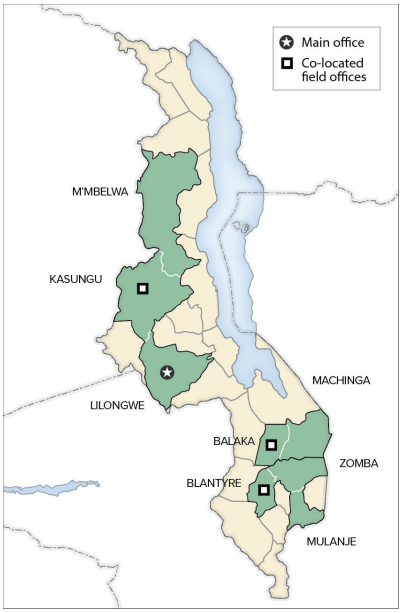

Figure 3: *Intervention Districts.* Main and Field offices refer to the offices of the USAID implementing partner.

Type	Project Component	Jul-17	Aug-17	Sep-17	Oct-17	Nov-17	Dec-17	Jan-18	Feb-18	Mar-18	Apr-18	May-18	Jun-18	Jul-18	Aug-18	Sep-18	Oct-18	Nov-18	Dec-18	Jan-19	Feb-19	Mar-19
Survey	Baseline Survey																					
Survey	Endline Survey																					
Monitoring	IPA Site Visits																					
Bottom-Up	Market Committee Elections																					
Bottom-Up	Market Committee Trainings																					
Bottom-Up	Kickoff Meetings																					
Bottom-Up	Revenue Information Push Messages																					
Bottom-Up	Grievance Reporting System																					
Bottom-Up	Construction of Infrastructure Projects																					
Bottom-Up	Handover of Infrastructure Projects																					
Top-Down	Mobile Money																					
Top-Down	Vendor Counting																					
Top-Down	Target Generation																					
Top-Down	Incentives for Tax Collectors and Markets																					

Figure 4: Timing of data collection and intervention components

decentralization.

B Explanation of Experimental Protocols/Interventions

This section presents a detailed intervention timeline and descriptions of each treatment component.

B.1 Bottom-Up Treatments

The first intervention bundle was designed to increase vendors' willingness to pay market taxes voluntarily. Above we identified three key reasons for low compliance: vendors feel that they receive little in return, especially with regards to market services; they don't see the government as accountable to vendors (citizens) and do not trust it to provide market services in the absence of stronger accountability; and they feel excluded from the tax system's structure. To address these barriers, we implemented a four-component intervention. Markets assigned to receive the bottom-up treatments received *all* of the components described below (except for Step 1: Facilitating Market Committee Elections, which only markets without valid market committees received). Our pilot research with vendors suggested that each component is unlikely to have a large impact on its own; the intervention needed to identify problems through the meetings, provide a costly signal of government commitment through the construction grants and then improve transparency to sustain any positive changes, and to empower vendors to monitor and sanction local officials for how they use market revenues.

Step 1: Facilitate Market Committee Elections

Not all markets in the Bottom-Up and Bottom-Up and Top-Down treatment groups had valid market committees. Invalid market committees are those that were formed without following council approved procedures — such as committees that were directly imposed on markets by the government — or whose terms have expired. As these market committees should serve to represent the market vendors' interests and interact with the government, it is important that vendors see these committees as legitimate extensions of their own interests. Only the markets in Mulanje district had valid market committees. As such, between October 2017 and December 2017, new elections were held in the 54 Bottom-Up treatment markets that did not have valid market committees. All newly elected market committees received a training in which committee members learned about on the proper organizational structure for the

committee and the roles and responsibilities of the market committees. These trainings, co-run by the District Councils' District Capacity Building staff and LGAP district staff, took place in December 2017 and January 2018 and underscored that regulations governing committees are government-sanctioned, to increase their validity and to signal to committee members that they themselves are legitimate actors in the governing structure. As the market committees played key roles in some of the other Bottom-Up interventions, vendors had to be able to see them as an interface between them and the government.

Step 2: Facilitate Meetings Between Vendors, Market Committees, and Local Government After ensuring that each treatment market had a valid market committee, the implementation partner facilitated public meetings to address vendors' exclusion from the taxation system. These meetings included vendors, market vendors' committees, and local government officials at treatment markets. The latter group included a representative from the District Finance Office, the market's tax collectors, the market/zone managers, the local ward councilor, and group village headmen. These meetings, which took place between January and February 2018, were observed by the LGAP district coordinator and included the following:

- 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.
- 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 the mHub SMS Market Revenue Reporting and Grievance Reporting systems[10]. 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 kickoff meetings, with a median of 44 registrations. Not all vendors registered for the system. The median proportion of vendors who attended who signed up for the system was 0.729.[11]

[10]See Step 4 for more information.

[11]These statistics exclude M'mbelwa markets, for which the implementing partner did not provide kickoff meeting reports.

- Documentation and discussion of the state of market services, including toilets, sanitation, security, and infrastructure. Vendors had the opportunity to develop a list of priorities – from an approved list of six project types[12] – for how the government should upgrade markets. This was done using a pairwise ranking method. This list was used to decide which infrastructure project a market would receive.[13]

This piece of the intervention resembles the classic case of tax bargaining, in which governments make policy concessions to vendors in return for tax revenues. We anticipated that this part of the intervention would increase citizens' sense that they have a say in how markets are run, alleviating the sense of exclusion that leads to low tax compliance. It may also have increased trust in local officials, especially if vendors viewed these meetings as "good faith gestures" from local government. In interviews done before the start of the project, market vendors in markets that had been visited by higher-up government officials always remembered and appreciated it. Finally, vendors received information about how revenues are used. A total of 3515 vendors attended these kickoff meetings in the 64 markets that received the Bottom-Up treatment bundle, with a median attendance of 61.5 vendors.[14]

Step 3: Jump Start Service Delivery in Markets

The meetings in Step 2 would be ineffective if vendors did not subsequently see improvements in services. Our scoping research indicated that increased revenues should be sufficient to maintain better services once infrastructure investments are made.[15] However, the condition of services in many markets at this point was so poor that even drastically improved tax compliance would be insufficient to fund necessary infrastructure improvements: low tax compliance leaves local government without funds for improvements, but citizens refuse to pay until services improve. This suggests the need to "jump-start" the tax-service provision loop by funding infrastructure improvements.

[12]Market shed, borehole, electricity, pathways, concrete slab, and refuse bins.

[13]See Step 3 for more information.

[14]These statistics exclude M'mbelwa markets, for which the implementing partner did not provide kickoff meeting reports.

[15]For example, we estimate that 10 days of fees from 10 vendors would cover the cost of a month of periodic trash pickup across several markets in an area, and that 25 days of fees from 10 vendors would cover the cost of a security guard for one market.

The initial meetings in Step 2 generated a list of priorities in each market. The implementing partner worked with the local government to implement priority projects in each market. Vendors were able to choose between a market shed, a borehole for water access, electricity, improved, formal pathways, concrete slab for stalls, and refuse bins. 46 markets chose a borehole – the other 18 chose a mix of the remaining project types. Each Bottom-Up market was allotted $5,000 for these infrastructure projects. These grants were clearly small and therefore not be sufficient to completely rehabilitate markets. They were designed, however, to be sufficient for small-scale infrastructure improvements. 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 and 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. We anticipated that this part of the intervention would serve as a costly signal of the government's commitment to improving service provision in markets.[16]

Step 4: Increase Transparency in the Taxation System via SMS Revenue Reporting and Grievance Reporting Systems

If local governments build new market infrastructure, but fail to maintain market services by providing ongoing sanitation and security services, tax compliance is unlikely to increase. To strengthen citizens' trust that their tax funds are being used well, and to facilitate bottom-up accountability between vendors and local government, citizens must also have information about government revenue and spending on an ongoing basis; a one-off meeting is unlikely

[16]Improved infrastructure may also have downstream effects on vendors' economic outcomes, as, in focus groups, customers cited poor sanitation as a reason for avoiding certain markets.

to yield long-term gains.

To improve citizens' access to information, we implemented an SMS messaging system designed to keep citizens informed about revenue collection and the type of expenditures made by district governments. 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.[17] At the meetings in Step 2, vendors were able to sign up for the SMS service. All vendors who signed up received the SMS messages unless they opted out.[18] Each subsequent month, vendors received a message with the amount of revenues the government raised from the market in the previous month, 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 information to vendors was centrally managed and required few steps. Messages were first sent out in January 2018,[19] and were last sent out in November 2018.

In addition to obtaining information on market revenues, vendors were able to use a related SMS system, also set up and managed by mHub, to report complaints and grievances about local government service delivery. Vendors were informed about this system in the meetings in Step 2, and markets were given materials to explain the system's use. 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.

This component of the bottom-up intervention was designed to improve transparency and information regarding how revenues are used. If revenues were being used well, this should

[17] `http://www.mhubmw.com/`

[18] An initial pilot of this idea in two markets demonstrated that vendors were eager to bear this small cost to sign up for this service.

[19] In markets where kickoff meetings had already taken place.

have helped to sustain high quasi-voluntary tax compliance by vendors. If, however, funds were not being used well, it may have had the opposite effect.[20] The grievance system was designed to give vendors more agency and enable them to make sure that revenues were used well.

Collectively, the four steps in the bottom-up intervention had the potential to significantly improve vendors' willingness to pay taxes. Overall, the intervention was designed to increase tax revenues; improve market services; increase vendor satisfaction with local governments; and empower vendors to hold officials accountable for how funds are spent. Increases in market services may have had additional benefits for vendors, including attracting more customers, increasing profits, and improving public health in and around markets.

B.2 Top-Down Treatments

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 reduce corruption on part of the tax collectors and market masters. Above we identified three key barriers to collection: inefficient collection systems; lack of knowledge of vendor numbers and expected revenues; and lack of incentives for tax collectors to work hard. All markets assigned to receive the top-down treatment received the following four components. Note that these components work together closely:

Step 1: Roll Out Mobile-Based Market Fee Payment System

To address the widespread potential for evasion, corruption, and inefficiency at the market level, markets in the top-down condition shifted to remitting market fees via mobile technology. Airtel Malawi was engaged to collect fees on behalf of the district council. Tax

[20] In our scoping and buy-in meetings, multiple stakeholders emphasized the importance of conveying the information about revenue clearly and simply, so that it would not be misunderstood by vendors and cause perverse effects, even in markets where the council is actually improving and contributing to market maintenance regularly.

collectors still collected fees from vendors and then gave the money to the market manager, who was then responsible for transferring the money to the Airtel agent[21] responsible for the market. The money was then transferred to the district council bank accounts. Airtel earned 2% of the fees as payment. Besides making payment of fees and their transfer to the district governments much more regulated and straightforward, this system was designed to allow the government to more reliably track how much each market collects in fees. It also made it easier to see if certain markets were not transferring fees 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 is the lack of a reliable estimate of anticipated revenue, which is required to determine collector benchmarks, monitor collector performance, and forecast local government revenue. The size of the market (measured in the number of vendors) changes over the course of the week, month, and year. Because of this, a formal registration system seemed cumbersome and likely to either marginalize irregular vendors or place an undue burden on them. Nevertheless, market size estimates are necessary to general revenue targets and forecast revenue. To address this issue, the implementing partner hired and trained vendor counters, who could not be market vendors or government staff. Counters visited each market at least four times a month — twice each on a market day and on a non-market day.[22] These vendor counters used a "walk-around" method — systematically walking through the market and recording 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 Num-

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

[22] Vendor counters visited markets about every two weeks. This means that some markets were visited five or six times in one month.

bers

The figures produced in Step 2 were used to determine 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 revenue collectors. Producing these revenue estimates was designed to serve multiple purposes. First, it should have allowed local governments to know how much revenue to expect. Second, it provided a way to evaluate the performance of a market, both in terms of vendors' tax compliance and tax collectors' ability to collect fees. For the latter group, it ideally provided a check against corruption and served as an incentive for better performance. 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

Under the current system, tax collectors lack incentives to enforce revenue collection. In some markets, tax collectors receive a fixed wage, with no incentives based on the revenues raised. In others, they receive commission pay.[23] However, among both groups tax collectors are paid less than $1 a day, and in vendor focus groups, vendors frequently noted that they believed that collectors were driven to bribery out of desperation and a need for supplementary income. Salaries are also often late, reducing the incentive to work hard.[24] Many tax collectors detailed the additional effort required to fully enforce tax compliance and punish the non-compliers; they hinted that they were possibly not as motivated as they could be, due to low incomes.

[23] All districts except for Machinga and Zomba had at least some tax collectors paid based on commission for at least some of the implementation period.

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

To address these issues, we implemented a bonus-type 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 collectors also received an individual incentive, which could have been a bicycle, fertilizer, certificates of excellence, mattresses, and 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.[25]

B.3 Explanation of Enforced Deviations from *Originally* Planned Interventions

Some of the interventions were significantly changed before roll out or at very nascent stages. This section details these changes.

- Incentives (TD):
  - As planned, if a market met its target, each tax collector was to receive an individual incentive, which could have been a bicycle, fertilizer, certificates of excellence, mattresses, and work suites. This was the case during the first month, during which markets received wheelbarrows, and tax collectors received individualized incentives such as mattresses or bicycles. After April, however, the district governments complained that these incentives were too individualized and stipulated that tax collectors' incentives had to be tied to the market. Markets that received incentives received a wheelbarrow and one bicycle, even if it had more than one tax collector. As such, the incentives were no longer tax collector-focused. These

[25] We note that the incentives were supposed to have been delivered to the markets and tax collectors who met their targets in the subsequent month. In actuality, the incentives were delivered between one and four months later. It is possible that this delay weakened the top-down treatment.

modifications to the design may have weakened the top-down treatment.

- Revenue Target Generation (TD):
  - As part of the intervention protocols, the research team developed a formula that took rainy season status, estimated compliance, and market and non-market day counts into account when developing revenue targets. However, district governments understandably pushed back on the complexity of this formula. This mean that the final revenue generation formula **did not** take into account whether the rainy season had started (which could depress the number of vendors present in the market), **nor** did it incorporate vendor counts from non-market days. As such, some of the nuance was sacrificed for ease of use.

C Deviations from Research Design and Intervention Plan

The project is described in its idealized form in Sections B.1 and B.2. In many markets, the project was indeed rolled out as intended. However, the rollout of the interventions was not as smooth in all districts and markets, with various interventions being delayed. In addition, initial struggles with the timing of the intervention delayed the project as a whole. We catalog the major issues in this section and address how we expect them to shape our analysis and influence our ability to detect results from the intervention.[26]

The issues that arose during this project are also informative in their own right. Implementing institutional reforms of any kind relies on a high degree of political buy-in and government capacity. All eight district governments, as well as the relevant central government ministries, agreed to the research design and all intervention components. We can

[26]There were many smaller issues as well. We only address the major issues here because it is more likely that these minor, market-specific issues were much more random in nature and as such could happen with any such project. In addition, due to their small-scale nature, they were more easily addressed. For example, Airtel agents in some markets had contentious relationships with the market managers when mobile money first started up. Airtel was able to work with is agents, and the district government with their market managers, ensuring that the tensions did not negatively affect money transfers.

therefore view deviations from the protocol not as unique to the Malawian context, but as representative of the types of issues that will arise in any attempt to build state capacity in similar low-income contexts where local governments have very limited baseline capacity. 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.

C.1 Project and Intervention Component Delays

The project as a whole was delayed several times. Initially scheduled to be completed in 2016 and 2017, bureaucratic hurdles meant that baseline data collection was not able to start until 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. Completing the project by summer 2018 was a major condition for the buy-in of the local governments because Malawian elections occurred in May 2019, and district governments did not want their hands to be tied by the randomized control trial nature of the experiment during the electoral campaign season. Finishing before the run-in to the elections was seen as preferable from our perspective as well, as the closer we came to the elections, the more likely it would that local politicians would be pressured to violate treatment assignment or that rival politicians in some of the districts and or wards would be tempted to start their own projects that were similar to this one. In both cases, our ability to find results from the project would have been compromised.

However, further issues with the implementing partner, the government of Malawi, and the district governments meant that although baseline data collection was completed by early September, many of the intervention components—especially the Top-Down treatments, which required more active buy-in from the district governments, and ended up being somewhat watered down—were not ready to be fielded until much later than originally

expected.

We pushed back endline data collection as late as we could without running into the next rainy season (which makes enumeration difficult), encroach on the peak campaign period of the May 2019 elections, or have endline data collection occur at a drastically different time of year than baseline data collection had.

As a result, most of the Top-Down treatments 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 the Top-Down treatments for approximately half of a year means that the treatments had less time to shape tax collectors' incentives and for the market staff to grow accustomed to the mobile money system, thus likely dampening the potential effect of the Top-Down treatments.

Within the Bottom-Up treatment arm, the infrastructure component was significantly delayed. As late as December 2017, the expectation was that construction would begin by April 2018 and would be finished by August 2018. However, the social and environmental reports and hydrologists' and scoping visits to the markets took longer than expected. In addition, some of the markets had chosen projects that were deemed too expensive, and so they had to be reconsulted, a process that took additional time. This meant that the implementing partner was not ready to approach contractors until July 2018. Hiring contractors once again took longer than expected, and construction did not actually begin until the very end of September. Only eight of the 64 projects started before October 30, when endline data collection began.

C.2 Treatment Implementation Issues

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

issues has the potential to weaken the interventions.

- Infrastructure (BU):
  - Because some markets chose projects as a first priority at their kickoff meetings that were eventually deemed too expensive, some markets did not receive their first choice infrastructure project. This issue affected about half of all Bottom-Up and Both markets, although the majority of these got their second choice.
  - Although hydrologists had visited the markets that had selected boreholes to assess where water could be found, no water was found after drilling in fourteen of the forty-six markets that were scheduled to receive a borehole. The implementing partner returned to these markets to offer alternatives: wheelbarrows and other cleaning supplies, or mobile refuse bins. Most markets received cleaning supplies.
  - As discussed above, construction was still ongoing during endline data collection. As such, some markets were visited for endline while construction was ongoing or before a handover ceremony—which were designed to formally connect the projects to the kickoff meetings and portray them as the result of tax compliance—had taken place.[27] 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.
- Mobile Money (TD):
  - The start of this component was significantly delayed due to issues with the Airtel system. Mulanje District was the last district to start using mobile money, not doing so until July 1, 2018.

[27] An additional concern is that many of these handovers did not take place on a market day, when the largest numbers of vendors could be in attendance, because of the schedules of the district officials who were supposed to be present.

- Standard Bank in Balaka District, contrary to the agreement made with the Balaka District Council, LGAP, and Airtel, deducted percentages of the fee transferred to the Balaka District Council's account at the bank. As a result, the Balaka District Council decided to suspend mobile money in its Top-Down markets on July 31, 2018, although some markets continued using mobile money for a few days.[28] Because of lack of cooperation from Standard Bank and the difficulty of transferring Balaka District Council's account to a different bank, Balaka markets did not resume using the mobile money system until the middle of October 2018.

- Vendor Counting:
  - This intervention component was initially one of the earliest top-down treatments to be rolled out, in January 2018. However, pushback from the district governments on the format of the vendor counting tool led to some intervention delays. We have vendor counts for January and February, but the finalized tool only began to be used in the second half of February 2018 and the beginning of March 2018 in some markets.

- Revenue Target Generation
  - DAI was not able to find records for any targets for Kasungu, Lilongwe, M'mbelwa, and Mulanje for May 2018, so we cannot be certain that Top-Down and Both markets in those districts actually received targets in that month.

- Incentives for Meeting Targets (TD):
  - In April, the first month that incentives were handed out, markets received wheelbarrows and tax collectors received individualized incentives such as mattresses, in the manner described above. However, the district governments decided that

[28] There were three Top-Down and three Bottom-Up and Top-Down markets in Balaka, representing 9.375% of the Top-Down markets.

these incentives were too individualized. They stipulated that the tax collector incentives should also be tied to the market and should be reusable by other individuals. As such, after the first month, all markets that met their targets received a wheelbarrow as a market-level incentive, and also received one bicycle. A market received only one bicycle, even it it had more than one tax collector. The next time that market met its target, it received another bicycle. These bicycles were to be usable by all market staff. This change undermines the intended intervention causal mechanism; incentives were designed to increase tax collector effort, but in order to accomplish this, they should have been targeted to individual tax collectors. The implementing partner did not inform us of this change until March 2019.

- Because market revenues for Balaka, Machinga, and Zomba for November 2018 did not become available until much later than expected, it was impossible to assess in December 2018 whether markets in these districts had met their November 2018 targets. Therefore, no markets in these districts received any incentives in December 2018.

- As discussed above, several months often elapsed between revenue collection and updating targets, and between meeting targets and receiving incentives. The length of these delays varied from district to district and from month to month.

- The implementing partner found in an audit after the intervention had been completed that a significant proportion that received incentives actually did not meet their targets, and that some markets that had met their targets did not receive incentives. They were not able to find a systematic explanation for why this occurred. It is possible that the delay and lack of congruence between incentives and market revenues weakened the top-down treatment.

C.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. Few of these protests were directly linked to the intervention components.[29] Often, these protests had to do with lack of adequate services in markets, serving to underline the importance of the infrastructure intervention, although this was not always the case. For example, Liwonde Central Market in Machinga—one of the largest markets there—sent no fees to the district council from February 2018 to December 2018 because vendors wanted to pressure the council to stop vendors from trading outside the bounds of the market. It is possible that these episodes of protest and boycott indicate that vendors in these markets are particularly aggrieved. This may make them harder to affect, in terms of increasing tax compliance, but also may make them more receptive to the infrastructure project component.

C.4 Competing Treatments

LGAP involves support for local governments and decentralization. The implementing partner was mandated to raise demand for better service delivery, with a particular goal to increase voter and civic education. As part of this aspect of their responsibilities, the implementing partner rolled out a project June to August 2018—when voter registration was ongoing—that aimed to inform Malawian citizens about the importance of local ward councilors and their responsibilities. This project took the form of a series of 10 roughly one and a half minute long radio messages that played nationally throughout Malawi.[30] These

[29] After the endline survey was fielded in Wendewende and Chimbalanga Markets, vendors there stopped paying fees. At first, the district government accused enumerators of telling vendors that they did not get all of the infrastructure projects that they should, and that the district government had stolen the money for these developments. However, a multilateral investigation and visit to these markets absolved the enumerators of responsibility and showed that the problem was much more general—vendors were unhappy with the state of services in their market.

[30] 6 different radio stations broadcast these messages. Five broadcast between June 1 and July 31, and the sixth broadcast between June 25 and August 25.

messages were broadly focused on educating citizens about decentralization and ward councilors. However, several of the messages did mention revenue collection and market fees. One stated that "local councils are supposed to inform us [citizens] how they are using revenue collected through market fees and other means." This matches the description of part of the SMS message component of the Bottom-Up treatments. While we do not believe that this treatment was particularly strong, it may have raised expectations among those who did hear the messages that more information would be provided (in bottom-up and bottom-up and top-down markets) or that this information would be provided in general (in control and top-down only markets, which did not received monthly revenue updates). As such, this project does compete, albeit weakly, with the bottom-up treatment.

D Data Collection Strategy

The data for the evaluation come from three sources. First, we implemented baseline and endline surveys to collect individual-level data from vendors, tax collectors, and local government officials. These surveys were carried out by our data collection partner, Innovations for Poverty Action (IPA). Second, LGAP provided ongoing information on the experiment's implementation and facilitated the transfer of vital data from the district governments. Finally, IPA also performed monitoring visits to each 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.

D.1 Baseline and Endline Surveys

D.1.1 Market Vendors Survey

In each of the 128 markets, the goal was to sample 100 vendors for a total of 12,800 interviews at both baseline and endline. Different individuals were surveyed at baseline and endline, unless the same individual was chosen by chance. At random, eighty of these vendors were given a 'short' (15-minute) version of the survey that primarily measures tax compliance and a handful of demographic variables. The remaining 20 respondents in each market received a 'long' (one-hour) version of the survey that includes more detailed data collection on demographics; economic, political, and social variables; and tax perceptions and payments.[31] Vendors were selected using a modified random walk. Enumerator teams of ten scoped out the general shape of the market and then divided the market into five sections for pairs of enumerators. Each pair then divided their section into two, planned out a path that would take them past all vendors in their section, and then determined a skip pattern that meant that they would interview ten vendors each. Markets were visited on their market days to ensure that the sample estimates reflected markets when the largest number of vendors were present.[32] Vendors 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).[33]

At baseline, 12,389 surveys were successfully completed. Not all markets had 100 vendors when they were visited. At endline, 12,370 surveys were successfully completed. Once again,

[31] The long version of the survey also included a conjoint survey experiment: the conjoint at baseline tested the factors that affect perceived willingness to comply with taxation, including a measure of time horizons; at endline, the conjoint tested vendor attitudes toward civil society organizations.

[32] We decided to do this as we considered it possible that vendors that do not come every day are either less likely to pay taxes consistently because they do not feel as connected to the market, or more likely to pay taxes, as they might feel less secure. Analysis of the baseline data showed that vendors who report working at a market more frequently were less likely to self-report paying taxes, although we were unable to determine the cause of this. Regardless, the baseline data showed that individuals who come less frequently are different from those who do, justifying this sampling strategy.

[33] At the time of data collection, the exchange rate was USD1=MWK720, so these compensation rates varied from USD0.28-USD0.83.

not all markets had 100 vendors. In addition, one market was mistakenly visited on a non-market day. We use this survey to test the main hypotheses.

D.1.2 Tax Collectors Survey

Up to seven tax collectors in each market took a 20-30 minute survey including 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.[34] Tax collector surveys also included a list experiment designed to elicit a measure of corruption within a market.

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.

D.1.3 District Council Survey

All elected ward councilors, as well as selected appointed district officials[35], were given a 20-30 minute survey at baseline and endline measuring: knowledge of tax law; awareness of tax roles and responsibilities; ability to forecast and track local taxes; awareness of tax compliance levels, and revenues collected, in each market in their jurisdiction; knowledge and perceptions of market services; frequency of interactions with each market; and perceived priorities for spending market tax revenues.

At baseline, 298 ward councilors and district government officials completed the survey.[36]

[34]Most markets have fewer than seven tax collectors. For time purposes, if a market has more than seven tax collectors, seven are chosen at random. At endline, seven tax collectors were interviewed in only one market. Another market saw nine interviewed. Although we do not know why protocol was broken, it might be that the supervisor had more time.

[35]IPA attempted to identify and contact all district officials involved with revenue collection, accounting, and finance.

[36]One hundred one ward councillors and 197 local government officials.

IPA attempted to contact 306 councilors and officials, representing a response rate of 97.4 percent. At endline, 352 ward councilors and district government officials were interviewed.[37] It is likely that a significant number of individuals were interviewed both at baseline and at endline, especially councilors, as there was no election in between, although there was some turnover among district government officials. Nevertheless, this survey data could be considered a panel.

This survey included some market-level outcomes but as district councilors likely have both control and treatment markets in their wards, these data are primarily used descriptively, as a manipulation check, and to improve our understanding of how the intervention has affected district governance and officials' incentives. We believe that this data is too noisy to use for analysis of our hypotheses. Although ward councilors were asked about all markets in their ward, district officials were—for the sake of time—not asked about specific markets in their district. In addition, we cannot ensure that ward councilors had the correct market in mind when answering market-specific questions.

D.2 Monitoring Data

Throughout the intervention period between baseline and endline data collection, we collected and received information that allowed us to monitor compliance with the interventions and track changes in market behavior and tax compliance between baseline and endline. These data were useful for us during the implementation phase, but are also crucial for our analysis, as they allow us to account for issues in implementation. 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

[37] One hundred eight ward councilors and 244 local government officials. The number at endline is larger because we were able to obtain a more comprehensive sampling frame for local government officials.

monitor implementation of the project.

D.2.1 Data Exchange

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.[38] LGAP also served as the intermediary for data from the district governments, including market revenues. These data are important beyond even monitoring the impact evaluation's roll-out; they allow us to assess how revenues change for each market throughout the intervention period and are thus a crucial part of the post-intervention analysis.

On a monthly basis, LGAP collected the following information at the *market level*, for each market:

- The estimated number of vendors. This information was only available for TD markets, in the form of the number of vendors counted by the vendor counters employed as part of Step 2 of the TD interventions;
- The number of tax collectors;
- The revenue targets (only for TD markets);
- Total revenues from fees for each month;
- Data on how market revenues have been allocated/spent. This is especially important in terms of maintenance spending. This information was only available on the sub-office level;[39]

[38] 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.

[39] Markets within districts are organized into sub-offices. Even after LGAP began their assistance to the district governments, including professionalization education and capacity building—a component of LGAP not evaluated here—spending was tracked most consistently at the sub-office level. The spending information was not well reconciled and had significant gaps, so full exploration of these data is not possible at this time.

- Details on LGAP impact evaluation intervention activities that have occurred in each market;
- Details on other LGAP intervention activities that have occurred within 10km of the market;
- Details on non-LGAP activities within 10km of the mark

In addition, LGAP provided information at the district level, including the following:

- Summary of intervention activities;
- Whether the rainy season has begun;
- If treatment assignment has been altered, or if the incorrect market received a certain intervention;
- Whether tax collectors or market managers are transferring to other markets;
- Whether large numbers of vendors have been noticed moving to sell in different markets; and
- Whether any markets belonging to different treatment groups have begun to share information about the different elements of the interventions.

D.2.2 Market Visits

IPA also carried out market visits throughout the intervention period. These visits 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.

Between November 2017 and July 2018, IPA visited each district approximately four times; once every two months). Each time, IPA representatives visited a random assortment of

markets across treatment groups in that district. This means that approximately 25 percent of the study markets were visited during every two-month period. Thus, over the course of the eight months, all 128 were scheduled to be visited.[40] These visits were unannounced, so that market staff could not prepare and 'dress-up' the market, and to avoid any changes in vendor and staff behavior that would endanger the accuracy of the observations. The visits took place on market days to maximize the number of vendors present and to observe the market at the height of its activity. The order in which markets in each district would be visited was randomly determined.

These market visits lasted roughly three hours. About two hours were devoted to observing the market and carrying out short interviews, while the last hour was reserved for a focus group discussion with a small group of market vendors. Each visit had the following format:

1. **Anonymous Walk (Duration 45 Min):** The anonymous walks allowed us to assess for ourselves how the intervention was progressing in each market without having information filtered through individuals who may have a vested interest in the project's success. IPA's Research Associate (RA) anonymously walked around the market to collect the following information:

    - Estimated number of vendors
    - Quality of toilets and other services, including the availability of water and the condition of market security
    - Evidence of recent changes/construction
    - Presence of posters for SMS campaign
    - Overall economic activity in the market

[40] In the end, we received the full data for only 123 markets. One market (Kasera) was not able to be visited. Four markets (Katenje, Mankhaka, Santhe, and Mafundeya) were visited during rainy season or on a non-mkt day, which impacted data collection and led to incomplete data.

Data entry was done via tablet using Survey CTO.

2. **Market Manager Interview (Duration 20 Min):** After the anonymous walk was complete, the RA informed the market manager of his presence and did a short interview with the market manager. The interview covered relations in the market between tax collectors, the manager, and the market committee, changes in the market—including any construction—, fee collection, revenue targets, and mobile money. It also addressed spillovers, particularly vendor and tax collectors moving to different markets. Interviews were programmed in tablets; the majority of items were close-ended, but some involved open-ended responses. Data entry was via Survey CTO.

3. **Market Committee Chairperson Interview (Duration 20 Min):** Next, the RA conducted an interview with the market committee chairperson (if absent, another member of the market committee leadership). This survey focused on similar topics to the market manager survey. The instrument was coded in tablets; the majority of items were close-ended, but some were open-ended responses. Data entry took place using Survey CTO.

4. **Tax Collector Interviews (Duration 20-60 Min):** The RA then performed one to three interviews with tax collectors, depending on how many tax collectors are working the market. These interviews once again covered similar topics to the one with the market manager, but also include questions on job satisfaction. The instrument was programmed in tablets using Survey CTO; most items were coded, but some involved open-ended responses.

5. **Focus Group Discussion (FGD) (Duration 45 Min):** Finally, the RA organized and moderated a focus group discussion with eight market vendors. After the RA had chosen a space for the FGD with the assistance of the market manager, he strategically chose a diverse set of participants to best represent the market, while at the same time selecting vendors qualified enough to contribute to the data gathering effort.

The discussion focused on documenting which intervention activities are occurring in that market, how vendors felt about those interventions, and whether vendors had noticed any change in fee collection / compliance. IPA provided additional personnel for translation in M'mbelwa and Machinga markets, as languages different from Chichewa—the language common to most of the other districts where this project was carried out—are common there (Chitumbka and Chiyao, respectively). IPA provided a token of appreciation to FGD participants.[41] IPA recorded the FGD and used the recording to produce notes following guidelines provided by the research team. Summary notes were recorded in Excel.

IPA was responsible for immediately alerting NORC of any observed violations of the research design (e.g., control markets that were receiving treatment or markets selected to a treatment not receiving it, intervention components that were being executed in a manner inconsistent with the intended design) revealed via market visits or that IPA became aware of through other means. This required that IPA's RA developed a good grasp of the evaluation design and was aware of the randomization results and general intervention timeline. Only three violations were reported—one in a TD market and two in control markets. These appear to be isolated, minor anomalies and should have no substantive implications on the main analysis.[42]

D.3 Measures

This section lays out all of the measures used in testing our hypotheses. All of the measures are drawn from the data sources discussed in Section D. Measures drawn from the vendor survey questions included on both the long and the short versions of the survey instrument

[41] The FGD moderator bought each FGD participant a snack or a soft drink of their choice.

[42] In one TD market, vendors reported receiving SMS messages with revenue information. In one control market, a market manager said that the old committee quit because LGAP had announced that there would be elections. Lastly, fee collectors from a nearby TD market were sent to a control market because the district council distrusted the management of the control market and believed that they were not remitting enough fees.

can be averaged up to the market level, to create market level estimates.[43] This is only possible for these questions because we otherwise have too few observations—20 instead of 100—per market to create reliable market-level estimates.

D.3.1 Main Outcomes

Table 9 presents the variables used to test hypotheses H1–H3, which correspond to the main hypotheses we laid out in Section 2.1.

Table 9: **Main Hypotheses Outcome Measures**

Hypothesis	**Aspect Assessed**	**Analysis Level**	**Variable**[44]	**Source**
H1	Self-Reported Tax Compliance	Individual and Market	_tc1a, _tc1b	Vendor Survey
H1	Perceived Group Tax Compliance	Individual and Market	*_tc3a, *_tc3b, *_tc3c	Vendor Survey
H1	Evidence of Receipt (Proxy for Tax Compliance)	Individual and Market	tc2_date (if date on most recent receipt is within a week/10 days)	Vendor Survey
H2	Market Fees Received by Government	Market	Market Fees Received by Government in November 2017 and November 2018, Standardized by Market Fee Amount	Government of Malawi, via Data Exchange
H3	Interaction of Treatments	Individual and Market	All of the Above	NA

The measures used in Table 9 are:

- tc1a: 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]

[43] We will indicate below which questions appeared only on the long version of the survey, which at most 20 vendors per market completed.

[44] For the sake of space, we have included only the survey item codes here and in subsequent tables. Please see the instruments for more information on the specific question text.

for each time you happened to be able to pay your K100 fee in the last 5 days.

- tc1b: [continued from previous] ...place a token here [indicate location] for each day you paid something, but less than K100.
- tc3a-tc3c: a similar coin exercise where respondents were given 10 tokens and allocated them between a) how many vendors pay every day; b) how many vendors pay some days, and c) how many vendors never pay the market tax.
- tc2date: Respondents were asked whether they could show the enumerator the last receipt they received for paying taxes. If the answer was yes, and the receipt was from the last 7 days (last 10 days as a robustness check), this was coded as "1".

Both baseline and endline surveys included two additional attempts to measure tax compliance that we decided not to use for analysis. The first was a list experiment, where the treatment list contained "Market Fees" in addition to the control list–Kerosene, New clothes, Batteries, Airtime. Analysis of the baseline data, however, showed that this was not a reliable measure, producing tax compliance estimates of greater than one and less than zero for some markets. The second was a series of questions that asked about whether vendors had seen or interacted with a fee collector in the past five days. Piloting before the baseline survey had indicated that this might be a useful way to get around social desirability bias. However, the baseline responses to these questions produced very high average tax compliance estimates between 88.5% and 98.1% depending on how the questions were used, which does not match up with the empirical reality.

D.3.2 Causal Mechanisms

We also collected data on intermediate outcomes to measure the mechanisms we believe the treatments affected. These intermediate outcomes are collected below in two tables, one each for the Bottom-Up and Top-Down interventions. As in the previous section, they are

hypothesis-specific and the source of the variables is also provided. The measures in this section are all drawn from the vendor and tax collector surveys. Because most of these questions appear only on the long version of the vendor survey or the tax collector survey–meaning that we will have only a small number of observations per market–these measures were not be aggregated up to the market level, unless indicated.

Table 10: **Bottom-Up Causal Mechanism Outcome Measures**

Hypothesis	**Aspect Assessed**	**Analysis Level**	**Variable**	**Source**
H4	Trust in District Government	Individual	tr1	Vendor Survey (Long)
H4	Trust in Ward Councilor	Individual	tr2	Vendor Survey (Long)
H5	Satisfaction with District Government, Indirect	Individual	tr9e, tr9g, tr9h[45]	Vendor Survey (Long)
H6*	Satisfaction with Services	Individual	ms1, ms3, ms4, ms5, ms6, ms10	Vendor Survey (Long)
H6	Perceived Amount of Spending on Services	Individual	tc2_10	Vendor Survey (Long)
H7	Tax Morale, Literature Measure	Individual	*_tc2_29_*	Vendor Survey
H7*	Paying Tax as Duty	Individual	tc2_4b	Vendor Survey (Long)
H7	Motivation for Paying Taxes	Individual	tc2_15_impact[46]	Vendor Survey (Long)

The variables referenced in starred rows of Table 10 (bottom-up mechanisms) are:

- tr1: In your opinion or based on what you have heard, would you say the district government is trustworthy?

[45]There were several more questions about government performance included in the survey, but the way the intervention was changed meant that we were simply not able to affect these outcomes. These questions are tr9f, tr9g, and tr9i.

[46]We expect that this measure will be very hard to move. At baseline, 73.9% of respondents reported that they paid taxed because "it is the right thing to do." There might also be high social desirability bias with this question.

- tr2: In your opinion or based on what you have heard,would you say your ward councilor for this market is trustworthy?
- ms1, ms3, ms4, ms5, ms6: questions regarding respondent's satisfaction with market services including clean water (ms1), garbage collection (ms3), pathways (ms4), stalls (ms5), and security (ms6).
- ms10: In general, how satisfied are you with the developments in THIS market provided by the district government?
- tc2_4b: Do you Strongly Agree, Somewhat agree, somewhat disagree, or strongly disagree with the following statement: Paying taxes is a duty of all citizens, even when you do not approve of how elected officials spend money.

Table 11: **Top-Down Causal Mechanism Outcome Measures**

Hypothesis	**Aspect Assessed**	**Analysis Level**	**Variable**	**Source**
H8*	Perceived Coercive Pressure	Individual	tc5a, tc5b, tc2_15b	Vendor Survey (Long)
H9*	Perception of Fees Not Reaching District Government	Individual	tc9[47]	Vendor Survey (Long)
H9	Difference in Reported Compliance and Fees Reaching District Government	Market	(Total Calculated from Self-Reported Compliance - Market Fees Received by Government) in November 2017 and November 2018	Vendor Survey & Government of Malawi via Data Exchange[48]
H9	Tax Collector Bribery	Individual	list_control, list_treatment[49]	Tax Collector Survey
H10	Hours Spent in Market	Individual	difference between e10 and e11	Tax Collector Survey
H10	Number of Vendors Visited Each Day	Individual	e12	Tax Collector Survey

The variables referenced in starred rows of Table 11 (top-down causal mechanisms) are:

- tc5a: Agree/Disagree (4-pt scale): If I wanted, I could refuse to pay my market fee.
- tc5b: Agree/Disagree (4-pt scale): If all vendors together decided to stop paying market fees, we could do so successfully.
- tc2_15b: Agree/Disagree (4-pt scale): I pay market fees because I'll get in trouble if I don't
- tc9: For every MK 1,000 fee collectors in this market collect, how much do you think reaches the district government?

D.3.3 Indirect Effect Outcomes

Table 12: **Indirect Effects Outcome Measures**

Hypothesis	**Aspect Assessed**	**Analysis Level**	**Variable**	**Source**
H11	Likelihood of Voting in Next Election	Individual	election3	Vendor Survey (Long)
H11	Contacting District Government with Complaint (Behavioral Experiment)	Individual	b6	Vendor Survey (Long)
H11	Willingness to Sign Petition to District Government (Anonymous/Identified)	Individual	behavioral/b1	Vendor Survey (Long)

E Survey Descriptive Statistics

E.1 Demographic Variables

Vendor Survey

Baseline:

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 13: 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

Endline:

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 14: 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

Tax Collector Survey

Baseline:

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 15: 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

Endline:

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 16: 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

E.2 Outcome Variables

Vendor Survey

Baseline:

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

Table 17: 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

Endline:

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

Table 18: 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

Tax Collector Survey

Baseline:

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 19: 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

Endline:

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

Table 20: 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

F Disaggregating Intervention Components

When considering whether or not to scale these intervention bundles, it is important to understand which of the components in each bundle is driving effects. In other words, would it be possible to achieve the effects we observe by only scaling a subset of the components? Obtaining a definitive answer to this question would require an impact evaluation with a full factorial design in which each component was randomized independently, and each market could receive any combination of the components. In the absence of this kind of evidence, we hesitate to speak to the contributions of each component. We are confident only about the effects of the components when bundled.

Nonetheless, we can draw on our endline data to provide suggestive evidence regarding the relative impact of various components within each bundle. In Tables 21 and 22 we present means by treatment group for a series of questions asked at endline about which intervention components the respondents recall.[50] These questions allow us to examine which components appear to be particularly memorable—and therefore possibly particularly effective (or ineffective)—for those involved in the study. While remembering a component is not the same as altering attitudes or behavior in response to a component, we assert that memorability is a compelling outcome as well.

Within the BU intervention bundle, there were four components: market committee elections and trainings; market kickoff meetings; the SMS transparency campaign; and market infrastructure projects. Table 21 indicates that nearly all vendors—regardless of treatment group—believe the market committee to be elected. However, as expected, vendors in the BU and BOTH groups appear more likely to report that the market committee was trained in the past year. Vendors in the BU and BOTH groups are also more likely to report a meeting between vendors and local government, although we note that fewer than one-third

[50] To avoid over-interpretation of these patterns, we refrain from conducting difference-in-means tests across treatment groups, and instead just make qualitative statements about easily observed patterns.

of vendors report such a meeting in any of the treatment groups, and that 14.8 percent of vendors in TD markets report such meetings. Interestingly, tax collectors appear to be more aware of these meetings; as shown in Table 22, in BU and BOTH markets, tax collectors are highly likely to report a meeting between the local government and vendors took place in the last year.

Among vendors who report the occurrence of such a meeting and that they attended it, we see those in the BU and BOTH treatment groups are likely to report vendors selected an infrastructure project at the meeting. Tables 23 and 24 respectively show that vendors in the BU and BOTH treatment groups are more likely to attribute responsibility for market construction to district government, and responsibility for funding market construction to USAID.

Returning to Table 21, vendors in the BU and BOTH treatment groups are more likely to report being aware of a way to find out how much revenue is collected from the market and receiving a SMS from the district about revenue in the past six months, though we note that both of these are reported at relatively low rates. Collectively, these patterns indicate that the market infrastructure projects are likely contributing to treatment effects in the BU group. The market kickoff meetings and the SMS transparency campaign could also be contributing to treatment effects, but this contribution is likely driven by a large effect on a minority of vendors rather than by a moderate effect on all vendors.

Within the TD intervention bundle, there were also four components: vendor counting; revenue targets for tax collectors; tax collector incentives for meeting targets; mobile money revenue transfers. In general, as shown in Table 22, data from the tax collectors survey indicates that the tax collectors recall most of the components of the TD intervention bundle. Tax collectors in the TD or BOTH markets are more likely to report vendor counting. However, they are not more likely to report being incentivized to meet targets for revenue collection: in fact, almost no tax collectors report being paid with incentives. Tax collectors

in TD and BOTH markets are more likely to report their managers transfer revenue to the district via mobile money, and they report more frequent transfers. Table 25 indicates that the modal frequency of transfers in the Control and BU markets is once a month, while in TD and BOTH markets it is every 3–4 days. Collectively, these patterns indicate that the vendor counting and mobile money components of the TD intervention were most salient for the tax collectors.

Table 21: Vendor Survey Treatment Component Question Averages, by Treatment Group

Question (Response Range)	Control	BU	TD	Both
Is there an active market committee in the market? (1/0)	0.863	0.897	0.855	0.936
Is the market committee elected? (1/0)	0.959	0.968	0.974	0.980
Has market committee received any training this past year? (1/0)	0.075	0.174	0.080	0.243
In the past year, have there been any meetings between local government and vendors in this market? (1/0)	0.090	0.205	0.148	0.233
At the meeting, did vendors select an infrastructure project for the government to build in the market? (1/0)	0.476	0.883	0.526	0.867
Is there a way vendors can find out how much the district is collecting from this market, or how the money is spent? (1/0)	0.013	0.053	0.019	0.083
In the past six months, did you ever receive an SMS message about market revenue? (1/0)	0.005	0.056	0.020	0.074

[51]Once a year = 1, Once every few months = 2, Once a month = 3, A few times a month (2-3 times per month) = 4, Once a week = 5, Every 5-6 days = 6, Every 3-4 days = 7, Every other day = 8, Every day = 9.

[52]There were 64 respondents from the Control group, 67 respondents from the BU group, 74 from the TD group, and 59 from the Both group.

Table 22: Tax Collector Survey[51]Treatment Component Question Averages, by Treatment Group

Question (Response Range)	**Control**	**BU**	**TD**	**Both**
In the past year, have there been any meetings between local government and vendors in this market? (1/0)	0.270	0.692	0.265	0.655
Were you paid for any of your work over the past year with incentives? (1/0)	0.016	0.000	0.000	0.000
Does your supervisor transfer market fee revenues to the district via mobile money? (1/0)	0.063	0.078	0.700	0.725
How often does your supervisor transfer market fee revenues to the district government? (1-9)[52]	2.734	3.106	5.431	5.661
In the past year, has anyone come and counted all vendors in the market? (1/0)	0.293	0.562	0.959	0.966

Table 23: Breakdown of Responses to 'Who is primarily responsible for construction or improvements [in this market]?' by Treatment Group[53]

Response Category	**Control**	**BU**	**TD**	**Both**
Market Manager	0.145	0.132	0.194	0.063
Ward Councilor	0.266	0.198	0.296	0.195
Other District Government	0.306	0.419	0.324	0.421
Member of Parliament	0.105	0.054	0.074	0.063
USAID/DAI/LGAP	0.040	0.108	0.009	0.189
Market Vendors/Market Committee	0.129	0.090	0.102	0.063
Traditional Leaders/Chiefs	0.008	0.000	0.000	0.006

[53]Only individuals who answered yes to the question asking whether there had been any construction in the market during the past year were asked this question.

[54]Only individuals who answered yes to the question asking whether there had been any construction in the market during the past year were asked this question.

Table 24: Breakdown of Responses to 'Who primarily funded this construction [in this market]?' by Treatment Group[54]

Response Category	**Control**	**BU**	**TD**	**Both**
Market Manager	0.048	0.062	0.073	0.020
Ward Councilor	0.143	0.089	0.115	0.108
Other District Government	0.429	0.479	0.458	0.459
Member of Parliament	0.067	0.055	0.062	0.041
USAID/DAI/LGAP	0.105	0.233	0.042	0.284
Market Vendors/Market Committee	0.200	0.082	0.250	0.088
Traditional Leaders/Chiefs	0.010	0.000	0.000	0.000

Table 25: Breakdown of Responses to 'How often does your supervisor transfer money collected from market fees to the district?' by Treatment Group

Response Category	**Control**	**BU**	**TD**	**Both**
Once a year	0.031	0.000	0.000	0.000
Once every few months	0.422	0.394	0.042	0.000
Once a month	0.469	0.424	0.250	0.237
A few times a month (2-3 times per month)	0.016	0.000	0.000	0.000
Once a week	0.016	0.091	0.125	0.153
Every 5-6 days	0.016	0.045	0.111	0.085
Every 3-4 days	0.031	0.045	0.472	0.525
Every other day	0.000	0.000	0.000	0.000
Every Day	0.000	0.000	0.000	0.000

G Investigating Mobile Money Treatment

This analysis has not been completed yet. We theorized that mobile money should be one of the strongest top-down treatments, which should have been detectable using the revenue data, we see no evidence of this. We plan to investigate why this was by looking at whether the shock of mobile money started induced an increase in revenue in the short term, if not the long term. We will use month-market-level regressions with monthly revenue as DV and TD as binary for mobile money operating in that month.

H Investigating Revenue Imbalance

H.1 Models

Table 26: Hypothesis 2 Results Table – DIM and DID

Panel A: DIM Models				
	Market Fee Units (MFU)	MFU NAs as 0s	Mkt F. per Tax Col.	MFU per Tax Col.
BU	795.109	946.471	14,734.150	100.729
	(538.564)	(604.933)	(13,509.370)	(105.335)
TD	1,094.672*	1,205.754*	18,045.800	135.859
	(526.460)	(580.517)	(13,098.720)	(102.133)
Both		327.210	1,633.386	31.393
		(601.760)	(13,437.050)	(104.771)
Observations	123	125	123	123
Adjusted R^2	0.144	0.147	0.258	0.292

Panel B: DID Models							
	MFU Nov. '17 to Nov. '18	MFU NAs as 0s Nov. '17 to Nov. '18	MFU NAs as 0s Dec. '17 to Nov. '18	MFU Nov. '17 − Nov. '18	Mkt F. per Tax Col. Nov. '17 − Nov. '18	MFU Dec. '17 − Nov. '18	Mkt F. per Tax Col. Dec. '17 − Nov. '18
BU	882.725	1,186.500*	1,182.255*	−94.445	−16,754.230	37.642	−10,823.350
	(509.243)	(529.973)	(506.502)	(366.193)	(19,598.200)	(314.216)	(11,713.630)
TD	1,422.100**	1,574.969**	1,301.838*	−154.090	−15,217.170	286.041	2,872.600
	(513.984)	(529.973)	(511.062)	(348.530)	(18,381.990)	(304.281)	(11,343.260)
Both		623.333	543.875	−159.338	−24,934.680	91.878	−7,882.518
		(529.973)	(506.502)	(363.802)	(19,187.280)	(310.060)	(11,558.690)
Endline:BU	−77.742	−203.690	−196.699				
	(711.816)	(755.658)	(705.398)				
Endline:TD	−362.961	−377.851	−48.130				
	(707.126)	(743.244)	(700.170)				
Endline:Both		−253.408	−173.554				
		(754.800)	(705.789)				
Observations	234	253	245	108	107	119	119
Adjusted R^2	0.245	0.267	0.276	−0.093	−0.149	0.002	0.085

Notes *p<0.05; **p<0.01; ***p<0.001
Models include block fixed-effects.

Table 27: Hypothesis 2 Results Table – DIM and DID, Logged Outcome Variable

Panel A: DIM Models			
	MFU (Logged)	Mkt F. per Tax Col. (Logged)	MFU per Tax Col. (Logged)
BU	0.318	0.234	0.234
	(0.277)	(0.212)	(0.212)
TD	0.507	0.199	0.199
	(0.268)	(0.206)	(0.206)
Both	0.107	0.051	0.051
	(0.275)	(0.211)	(0.211)
Observations	123	123	123
Adjusted R^2	0.346	0.328	0.378

Panel B: DID Models					
	MFU Nov. '17 to Nov. '18 (Logged)	MFU Nov. '17 − Nov. '18 (Logged)	Mkt F. per Tax Col. Nov. '17 − Nov. '18 (Logged)	MFU Dec. '17 − Nov. '18 (Logged)	Mkt F. per Tax Col. Dec. '17 − Nov. '18 (Logged)
BU	0.905***	−0.470*	−0.502	−0.361	−0.367
	(0.269)	(0.235)	(0.253)	(0.212)	(0.217)
TD	0.967***	−0.285	−0.380	−0.209	−0.272
	(0.270)	(0.223)	(0.238)	(0.205)	(0.210)
Both	0.688*	−0.515*	−0.563*	−0.193	−0.237
	(0.269)	(0.233)	(0.248)	(0.209)	(0.214)
Endline:BU	−0.611				
	(0.367)				
Endline:TD	−0.482				
	(0.365)				
Endline:Both	−0.602				
	(0.367)				
Observations	234	108	107	119	119
Adjusted R^2	0.389	0.096	0.077	0.180	0.157

Notes *p<0.05; **p<0.01; ***p<0.001
Models include block fixed-effects.

H.2 Exploration of Pre-treatment Imbalance

The models in Appendix H.1 indicate that there are significant differences in market revenues between treatment groups at endline. However, the effects do not survive a difference-in-difference analysis. Tables 28 and 29 shed some light on why this may be. Table 28 shows that there were already significant differences between the treatment groups in November 2018, right at the beginning of the intervention. There are three caveats: first, it is certainly possible that the data we have are flawed, as we did not receive them until May 2017; second, we are missing revenue information for November 2017 for 17 markets; and third, some of the interventions had already started in November 2017. We must stress that the receipt measure would seem to indicate that tax payment went up at least in bottom-up markets (or tax payers' belief in the system went up, causing them to be more likely to request receipts?).

Table 28: November 2017 Market Revenue, By Treatment Group

Variable	Control[3]	BU[2]	Both[1]	TD[4]
Market Fees Collected (Kwacha)	113526[2,4]	258570.7[3]	176139.7	342885.7[3]
Market Fee Units Collected	824.333[2,4]	2019.874[3]	1398.448	2535.976[3]
Market Fees per Fee Collector	32752.77[1,2,4]	63294.42[3]	55116.04[3]	68432.33[3]
Market Fee Units per Fee Collector	263.429[2]	491.075[3]	456.21	560.692
Logged Market Fees Collected	10.991[2,4]	11.901[3]	11.578	11.956[3]
Logged Market Fee Units Collected	6.145[2,4]	7.027[3]	6.717	7.087[3]
Logged Market Fees per Fee Collector	10.091[2,4]	10.782[3]	10.549	10.682[3]
Logged Market Fee Units per Fee Collector	5.245[2,4]	5.897[3]	5.688	5.813[3]

At the same time, we do have some evidence that our groups were different despite randomization. Table 29 shows a series of pre-treatment market characteristics by treatment group. These data were obtained from LGAP's scoping exercise before the intervention had begun. We are absolutely certain that these were collected before treatment assignment. They indicate that top-down markets seem to be bigger than those in other treatment groups, with more vendors on average—both during the dry season and the rainy season—and more

tax collectors on average. That said, we stress that these differences are **not** statistically significant (although this may be a power issue). In addition, these data are very noisy and rough.

Table 29: Pre-Treatment Market Characteristics, By Treatment Group

Variable	Control[3]	BU[2]	TD[4]	Both[1]
Number of Tax Collectors (from Data Exchange)	3.438	3.375	4.938	3.344
Number of Tax Collectors	1.875	1.594	3.094	1.906
Number of Revenue Collectors	1.062	0.781	0.812	1.188
All Collection Personnel	3.25	2.375	3.906	3.25
Average Daily Number of Vendors (Dry Season)	178.661	228.393	357.63	229.701
Average Daily Number of Vendors (Wet Season)	106.089	143.777	207.309	130.763

Finally, it is possible that some external factor has influenced all revenue collection. Table 30 indicates that before the interventions began, more than the majority of tax collectors reported their supervisors transferring money to the district government once every few months, with the rest saying once a year. Comparing this table to Table 25, however, shows that tax collectors noticed their supervisors transferring money to the district government more often than before treatments began in *all* markets, even in control markets. It is possible that once district government realized the intervention was starting, they pushed all markets to transfer money more often, regardless of treatment status. This may have decreased the possibility of corruption.

Table 30: Breakdown of Responses to 'How often does your supervisor transfer money collected from market fees to the district?' by Treatment Group, at Baseline

Response Category	**Control**	**BU**	**TD**	**Both**
Once a year	0.294	0.356	0.259	0.270
Once every few months	0.676	0.630	0.667	0.689
Once a month	0.000	0.000	0.037	0.014
A few times a month (2-3 times per month)	0.029	0.014	0.012	0.000
Once a week	0.000	0.000	0.025	0.027
Every 5-6 days	0.000	0.000	0.000	0.000
Every 3-4 days	0.000	0.000	0.000	0.000
Every other day	0.000	0.000	0.000	0.000
Every Day	0.000	0.000	0.000	0.000

I 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, kickoff 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. In lieu of adjudicating between these explanations, we present excerpts from the qualitative data about the real-time experiences and perceptions of vendors and tax collectors collected during the monitoring period. We emphasize that we have only processed and analyzed a fraction of these data: the 125 focus group discussions yielded over 375 pages of single-spaced text, and the more than 500 interviews yielded thousands of pages of text. We provide indicative quotes from the vendor focus group discussions in the BOTH markets to support the plausible explanations presented above:

> *"I have been expecting to see [the development] the people have been talking [about] but there is nothing, so when paying the fees I become angry since it's like wasting my money on something that I don't understand."*
>
> *"Most [of the] relationship between the local government and the [market] committee is about the fee collection, not the good of the market. Like when the local government come here they talk about the fees collection only, not the welfare of*

the vendors."

"It hurts when we heard that this market makes a lot of money but [the local government] can't support it when there is a need."

"We have been paying in the market daily and there is a lot of money collected here but the district is failing to fix problems in this market."

"There are no infrastructures here, there are no good toilets here, but yet we have been paying and we have been told that this money will be used here."

J Spillovers

J.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).

J.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."[55] 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

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

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.[56] 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.

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

[56] 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.

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 **??**, 32, and 33.

Table 31: 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.

J.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

Table 32: 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 33: 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.

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 34: 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 35: 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 36: 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.

K 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 the following problems:

1. Endline data collection occurred before mobilization for its infrastructure project had started.
2. A vendor from this market sent in a grievance message that was not responded to.
3. The 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 had initially considered designating markets that did not get their first choice infrastructure project as a compliance issue, but there was not enough variation here (more than 70% of markets did not get their first choice project).

We say that a top down treatment market has a compliance issue if it had one of the following problems:

1. A market met its target but did not receive an incentive.
2. The incentive a market was supposed to receive arrived delayed.
3. Mobile money was not active in any given month after May 2018.

We had initially considered using whether vendor counting was delayed, and whether we were unsure whether a revenue target was communicated each month as compliance measures, but

there was not enough variation for either variable. Only 4.69% saw vendor counting delayed, and 71.9% saw a month in which it was unclear whether revenue targets were relayed. For the latter, this was June 2018, when the implementing partner was not able to verify that targets had gone out to any markets in 5 districts.

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.

We present the results for models where we interact treatment and where we treat treatment groups as separate.

K.1 Treatment Arms Interaction

Table 37: Compliance IV Regression 2nd-Stage Factorial Approach

	Dependent variable:					
	Self-Rep. Compl		Group-Per. Compl.		Recent Rcpt.	
	Strict	Relaxed	Strict	Relaxed	Strict	Relaxed
BU Treat. - Str.	0.320		0.566		0.296*	
	(0.246)		(0.400)		(0.126)	
TD Treat. - Str.	0.407		0.131		0.198	
	(0.233)		(0.332)		(0.103)	
BU Treat. - Rel.		0.130		0.214		0.112**
		(0.086)		(0.144)		(0.036)
TD Treat. - Rel.		0.179*		0.056		0.083*
		(0.085)		(0.126)		(0.034)
BU Treat. - Str. * TD Treat. - Str.	−1.929		−1.593		−0.974	
	(1.231)		(1.824)		(0.527)	
BU Treat. - Rel. * TD Treat. - Rel.		−0.297		−0.219		−0.145**
		(0.152)		(0.246)		(0.056)
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.

K.2 Treatment Groups

Table 38: 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.

L General Robustness Models

L.1 Main Outcomes

L.1.1 Other Main Outcome Specifications

Table 39: 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 40: H1: Group-Perceived Tax Compliance Robustness Models

	Group-Perception of Always Complying			
	Market DIM	Market DID	Individual DID	Individual DID 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 41: 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.

L.1.2 Interacting BU and TD Treatment Assignment

Table 42: Analysis as Factorial Design (w/ Int., no Int.)

	Self-Reported Full Compliance		Self-Reported Always Complying		Evidence of Receipt 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.

L.1.3 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). In the models presented in this section, only completely invalid entries (greater than 5 or 10 or less than 0) are treated as NAs, and *all 0* outcomes are retained.

Table 43: H1: Self-Reported and Group-Perceived Tax Compliance (Incl. All 0s)

	Self-Reported Full Compliance			Perception of Others' Always Complying		
	Individual DIM	Market DIM	Market DID	Individual DIM	Market DIM	Market DID
BU	0.158	0.149	0.156	0.179	0.195	0.218
	(0.101)	(0.118)	(0.115)	(0.142)	(0.169)	(0.156)
TD	0.178**	0.177	0.057	0.050	0.049	0.103
	(0.081)	(0.118)	(0.115)	(0.118)	(0.169)	(0.156)
Both	0.040	0.044	0.058	0.033	0.084	0.158
	(0.110)	(0.118)	(0.115)	(0.153)	(0.169)	(0.156)
Endline:BU			−0.007			−0.023
			(0.163)			(0.221)
Endline:TD			0.119			−0.054
			(0.163)			(0.221)
Endline:Both			−0.014			−0.074
			(0.163)			(0.221)
Observations	12,360	128	256	12,319	128	256
Adjusted R^2	0.064	0.348	0.261	0.116	0.114	0.222

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.

L.1.4 Alternative Outcomes

Table 44: H1: Partial Self-Reported Tax Compliance – Individual and Market Level

	Self-Reported Partial Tax Compliance Os as NA			Self-Reported Partial Tax Compliance Os as 0s		
	Individual DIM	Market DIM	Market DID	Individual DIM	Market DIM	Market DID
BU	−0.040	−0.040	0.022	−0.036	−0.034	0.022
	(0.037)	(0.047)	(0.046)	(0.036)	(0.046)	(0.046)
TD	−0.034	−0.049	−0.041	−0.033	−0.052	−0.041
	(0.042)	(0.047)	(0.046)	(0.040)	(0.046)	(0.046)
Both	−0.072**	−0.091*	0.007	−0.072**	−0.087*	0.007
	(0.035)	(0.047)	(0.046)	(0.034)	(0.046)	(0.046)
Endline:BU			−0.062			−0.056
			(0.066)			(0.065)
Endline:TD			−0.008			−0.011
			(0.066)			(0.065)
Endline:Both			−0.098			−0.093
			(0.066)			(0.065)
Observations	11,823	128	256	12,362	128	256
Adjusted R^2	0.066	0.277	0.407	0.065	0.251	0.394

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 45: H1: Perception of Other Vendors Paying Fee Sometimes – Individual and Market Level

	Perception of Others' Sometimes Complying 0s as NA			Perception of Others' Sometimes Complying 0s as 0s		
	Individual DIM	Market DIM	Market DID	Individual DIM	Market DIM	Market DID
BU	−0.191***	−0.239***	−0.170*	−0.195***	−0.241***	−0.170*
	(0.074)	(0.090)	(0.094)	(0.075)	(0.091)	(0.095)
TD	−0.022	−0.039	−0.202**	−0.022	−0.038	−0.202**
	(0.077)	(0.090)	(0.094)	(0.078)	(0.091)	(0.095)
Both	−0.154**	−0.196**	−0.123	−0.163**	−0.203**	−0.123
	(0.076)	(0.090)	(0.094)	(0.077)	(0.091)	(0.095)
Endline:BU			−0.069			−0.071
			(0.133)			(0.134)
Endline:TD			0.163			0.164
			(0.133)			(0.134)
Endline:Both			−0.072			−0.079
			(0.133)			(0.134)
Observations	12,295	128	256	12,320	128	256
Adjusted R^2	0.139	0.102	0.274	0.139	0.106	0.270

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 46: H1: Perception of Other Vendors Never Paying – Individual and Market Level

	Perception of Others' Never Complying 0s as NAs			Perception of Others' Never Complying 0s as 0s		
	Individual DIM	Market DIM	Market DID	Individual DIM	Market DIM	Market DID
BU	−0.005	0.036	−0.056	−0.009	0.014	−0.056
	(0.127)	(0.173)	(0.154)	(0.121)	(0.155)	(0.144)
TD	−0.026	−0.005	0.101	−0.026	−0.005	0.101
	(0.086)	(0.173)	(0.154)	(0.083)	(0.155)	(0.144)
Both	0.090	0.089	−0.030	0.080	0.069	−0.030
	(0.134)	(0.173)	(0.154)	(0.126)	(0.155)	(0.144)
Endline:BU			0.093			0.071
			(0.218)			(0.203)
Endline:TD			−0.106			−0.106
			(0.218)			(0.203)
Endline:Both			0.119			0.099
			(0.218)			(0.203)
Observations	12,292	128	256	12,317	128	256
Adjusted R^2	0.090	0.047	0.070	0.088	0.056	0.080

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 47: 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 48: H1: Outcome 3 - Tax Collector Rarely or Never Does **Not** Give You A Receipt When You Pay Fee

	No Receipt When Paying		
	EL DIM	BL-EL DID	EL DID (Lagged DV)
Endline		0.052***	
		(0.020)	
BU	−0.059***	−0.014	−0.059***
	(0.020)	(0.016)	(0.020)
TD	−0.004	0.010	−0.003
	(0.019)	(0.017)	(0.019)
Both	−0.054***	−0.017	−0.055***
	(0.019)	(0.017)	(0.019)
Endline:BU		−0.036	
		(0.032)	
Endline:TD		−0.020	
		(0.030)	
Endline:Both		−0.037	
		(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
			All models have individuals as unit of analysis.
			All include block fixed-effects.
			Endline only models include enumerator fixed-effects as well.
			All models have SEs clustered on market.
			Lagged DV model includes baseline market average of DV.
			Outcome is on binary.

L.2 Intermediate Outcomes

L.2.1 BU Outcomes

Table 49: 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 50: 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 51: 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 52: 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 53: 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.
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.

L.2.2 TD Outcomes

Table 54: 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 55: 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 56: 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 hours in a day. Outcome 2 is a positive integer.

L.2.3 Binary Versions of Significant Intermediate Outcomes

Table 57: Trust in Ward Councilor and Treatment Status Cross-Tab, Raw Numbers

	BU	BU & TD	Control	TD
Not at all trustworthy	89	99	128	139
Not very trustworthy	144	140	126	147
Somewhat trustworthy	243	249	247	207
Very trustworthy	145	127	110	107

Table 58: Trust in Ward Councilor and Treatment Status Cross-Tab, Percentages

	BU	BU & TD	Control	TD
Not at all trustworthy	0.143	0.161	0.209	0.232
Not very trustworthy	0.232	0.228	0.206	0.245
Somewhat trustworthy	0.391	0.405	0.404	0.345
Very trustworthy	0.233	0.207	0.180	0.178

Table 59: Ward Councilor is Trustworthy and Treatment Status Cross-Tab, Raw Numbers

	BU	BU & TD	Control	TD
0	233	239	254	286
1	388	376	357	314

Table 60: Ward Councilor is Trustworthy and Treatment Status Cross-Tab, Percentages

	BU	BU & TD	Control	TD
0	0.375	0.389	0.416	0.477
1	0.625	0.611	0.584	0.523

Table 61: Agree Paying Tax is Duty (4-Point Scale) and Treatment Status Cross-Tab, Raw Numbers

	BU	BU & TD	Control	TD
Strongly Disagree	12	15	20	14
Somewhat Disagree	21	24	24	21
Somewhat Agree	96	112	120	119
Strongly Agree	509	490	466	468

Table 62: Agree Paying Tax is Duty (4-Point Scale) and Treatment Status Cross-Tab, Percentages

	BU	BU & TD	Control	TD
Strongly Disagree	0.019	0.023	0.032	0.023
Somewhat Disagree	0.033	0.037	0.038	0.034
Somewhat Agree	0.150	0.175	0.190	0.191
Strongly Agree	0.798	0.764	0.740	0.752

Table 63: Agree Paying Tax is Duty and Treatment Status Cross-Tab, Raw Numbers

	BU	BU & TD	Control	TD
0	33	39	44	35
1	605	602	586	587

Table 64: Agree Paying Tax is Duty and Treatment Status Cross-Tab, Percentages

	BU	BU & TD	Control	TD
0	0.052	0.061	0.070	0.056
1	0.948	0.939	0.930	0.944

Table 65: Causal Mechanism Outcomes, Binary Variable Versions (Endline DIM)

	Dependent variable:					
	Trust Local Gov.	Trust Ward Counc.	Satisfied w/ Services	Satisfied w/ Water Access	Paying Tax is Duty	Pay Because Consequences
	OLS	*OLS*	*OLS*	*OLS*	*OLS*	*OLS*
BU	0.076**	0.045	0.112**	0.225***	0.015	0.012
	(0.028)	(0.032)	(0.037)	(0.060)	(0.012)	(0.012)
TD	0.026	−0.072*	0.038	0.049	0.014	0.028**
	(0.032)	(0.028)	(0.034)	(0.050)	(0.012)	(0.010)
Both	0.070*	0.027	0.070	0.097	0.010	0.023*
	(0.028)	(0.031)	(0.037)	(0.056)	(0.012)	(0.011)
Observations	2,509	2,447	12,365	2,517	2,531	2,518
Adjusted R^2	0.151	0.117	0.132	0.114	0.038	0.225

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

This version of the endline difference-in-means analysis simplifies interpretation of the effect of the treatments on important intermediate outcomes. It is important to note, however, that these results *accompany*, but do not replace or invalidate those from the models where the outcomes are treated as on a four-point scale.

The first model in Table 65 confirms the positive impact of the Bottom-Up treatment on vendors' trust in the local government. Vendors in the BU treatment condition were 7.6% more likely to say that they found the local government at least somewhat trustworthy than those in control markets at endline; vendors in the Both treatment condition were 7.0% more likely. When it comes to trust in the ward councilor, however, treating the outcome as a binary variable does paint a somewhat different picture from the main analyses. At endline, vendors in the Top-Down treatment condition were 7.2% less likely to say that their ward councilor was at least somewhat trustworthy. The first set of four cross-tabs above help illuminate why. As the second set of tables shows, the proportion of vendors overall selecting "Not at all trustworthy" or "Not very trustworthy" is similar in the Control, Bottom-Up, and Both markets, and is very different in Top-Down markets. However, as we can see in the first two tables, in the Bottom-Up and Both markets, more vendors picked the higher choice *within* the not trustworthy and trustworthy halves than in Control markets. Treating outcome as binary obfuscates this difference.

The results of the satisfaction outcomes show that, at endline, vendors in Bottom-Up only markets were 11.2% more likely to report being at least somewhat satisfied with their services and 22.5% more likely to state that they were at least somewhat satisfied with their access to water, compared to vendors in Control markets. This underscores the power of the Bottom-Up treatment, but also emphasizes that something occurred in Both markets that undercut the treatment.

Unlike in the main analysis, the treatments did not seemingly impact whether vendors agreed at least somewhat that paying tax is a duty. The second set of four cross-tabs in this section

above once again help us understand why this is: the difference detected by the main analysis comes from the greater proportion of vendors who report that they "Strongly Agree" in the Bottom-Up markets compared to the Control group vendors.

On the other hand, vendors in Top-Down markets at endline were 2.8% more likely to agree at least somewhat that they pay market fees because there will be consequences if they do not, compared to those in Control markets. Vendors in Both markets were 2.3% more likely. This confirms the main analyses and seems to indicate that enforcement, or the perception of enforcement, did increase in markets that received the top-down interventions.

M Explanation of Deviations from PAP

M.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.

N Analyses Still In Progress

The following pre-specified analyses are still in progress:

- Multiple comparisons correction
- IV approach (compliance analysis)
- More in-depth qualitative (FGD, monitoring interviews, etc)
- Subgroup analyses
- Multilevel modeling
    - Mostly done; results seem similar