Tax Farming Redux:
Experimental Evidence on Performance Pay for Tax Collectors

Adnan Q. Khan, Asim I. Khwaja, and Benjamin A. Olken*

June 2015

Abstract

Performance pay for tax collectors has the potential to raise revenues, but might come at a cost if it increases the bargaining power of tax collectors vis-a-vis taxpayers. We report the first large-scale field experiment on these issues, where we experimentally allocated 482 property tax units in Punjab, Pakistan into one of three performance-pay schemes or a control. After two years, incentivized units had 9.4 log points higher revenue than controls, which translates to a 46 percent higher growth rate. The scheme that rewarded purely on revenue did best, increasing revenue by 12.9 log points (64 percent higher growth rate), with little penalty for customer satisfaction and assessment accuracy compared to the two other schemes that explicitly also rewarded these dimensions. The revenue gains accrue from a small number of properties becoming taxed at their true value, which is substantially more than they had been taxed at previously. The majority of properties in incentivized areas in fact pay no more taxes, but instead report higher bribes. The results are consistent with a collusive setting in which performance pay increases collectors' bargaining power over taxpayers, who either have to pay higher bribes to avoid being reassessed, or pay substantially higher taxes if collusion breaks down.

*This project is the result of collaboration among many people. We thank the editor, Lawrence Katz, the second editor, Andrei Shleifer, and four anonymous referees for helpful comments. We thank Jon Hill, Donghee Jo, Alyssa Lawther, Kunal Mangal, Wayne Sandholtz, Mahvish Shaukat, Gabriel Tourek, He Yang, and Gabriel Zucker for outstanding research assistance in Cambridge and Zahir Ali, Osman Haq, Turab Hassan, Zahra Mansoor, Obeid Rahman, Shahrukh Raja, Adeel Shafqat, and Sadaqat Shah for outstanding research assistance in Lahore. We thank all the Secretaries, Director Generals, Directors, the two Project Directors from the Punjab Department of Excise and Taxation, the Punjab Finance, Planning and Development departments and the Chief Secretary and Chief Minister's offices for their support over the many years of this project. Financial support for the evaluation came from 3ie, the IGC, and the NSF (under grant SES-1124134), and financial support for the incentive payments described here came from the Government of the Punjab, Pakistan. The views expressed here are those of the authors and do not necessarily reflect those of the many individuals or organizations acknowledged here.

1 Introduction

Tax systems throughout the developing world collect substantially less revenue as a share of GDP than their counterparts in developed countries (Gordon and Li 2009, Kleven et al. 2014). While there are many sources of differences, an important one is the substantial role played by the tax officials in assessing, enforcing, and auditing taxes. Combined with relatively low wages and limited performance rewards, the temptations for tax inspectors to collude with taxpayers to reduce tax receipts are great.

One possible solution to this problem is to tie the compensation of tax staff to the revenue they generate. This is an old idea: historically, states from the Roman Empire through the French Monarchy (Bartlett 1994, White 2004) sold the rights to collect taxes to "tax farmers," who then kept a fraction (or in some cases all) of the tax revenue they collected from a given area. US states similarly experimented with highly incentivized "tax ferrets" to collect property taxes in the 19th century. Though tax officials in these historical regimes were unpopular, and the world has largely moved to salaried tax officials (Parrillo 2013), countries such as Brazil, Peru, Pakistan, and others have begun to reconsider incentives for tax staff (Das-Gupta and Mookherjee 1998, Kahn et al. 2001) as a way to improve tax compliance.

The challenge, however, is that by strengthening the bargaining ability of tax collectors, performance incentives may not only lead to taxpayer dissatisfaction, but may also alter the division of rents from collusion without necessarily increasing revenue raised by the government. To see this, consider a simple bargaining setting in which a tax collector colludes with a taxpayer to reduce the tax assessment in exchange for a bribe. If there is no cost to either party from reducing tax liability, then performance pay for tax collectors will simply raise the bribe paid with no impact on revenue, as the taxpayer now has to compensate the collector's foregone incentive payment with a higher bribe. In more realistic settings, where there is some cost to either party from reduced tax liabilities, there will be two different effects: some taxpayers will continue in the collusive, low-tax equilibrium but pay higher bribes, while others will end up paying higher taxes and lower bribes as they switch from the collusive, low-tax equilibrium to a non-collusive, higher-tax equilibrium. Performance pay could thus have heterogeneous effects on tax revenue and bribes among taxpayers. Whether performance pay actually leads to increased revenue – and at what cost in terms of higher bribes and potentially foregone taxpayer satisfaction – is therefore ultimately an empirical question.

In this paper, we provide what is to the best of our knowledge the first experimental evidence on these questions. Working with the Punjab, Pakistan provincial government, we randomly allocated tax officials in the entire provincial urban property tax department, which consists of 482 property tax units (known as circles), into one of three versions of performance-based pay schemes or a control group. A total of 218 circles, consisting of about 550 tax personnel, were randomly allocated to one of the three treatment groups, for two fiscal years. The incentives were large: the three-person tax team in each treated circle was collectively given an average of 30 percent of all tax revenues it collected above a historically-predicted benchmark. Many personnel in treated areas were able to double their baseline salaries or more through these incentives.

Given concerns about potential negative impacts of high-powered incentives, the three schemes varied in both the extent to which they based performance pay explicitly on taxpayer satisfaction and accuracy of assessment in addition to revenue, and the extent to which they allowed for subjective evaluation on the part of the tax department. The "Revenue" scheme provided incentives based solely on revenue collected above a benchmark predicted from historical data. To address multi-tasking concerns and in an effort to incentivize accurate assessment (and hence also tackle collusion) (Holmstrom and Milgrom 1991), the "Revenue Plus" scheme provided incentives exactly as in the Revenue scheme, but made adjustments (plus/minus three-fourths of baseline salary) based on whether the circle ranked in the top, middle, or bottom third of circles in terms of taxpayer satisfaction and accuracy of tax assessments, as determined by an independent survey of taxpayers. To allow for more subjective assessments rather than purely formulaic criteria (Baker et al. 1994, MacLeod 2003), the third scheme, "Flexible Bonus," took this a step further by both rewarding collectors for a much wider set of pre-specified criteria set by the tax department, and by allowing for subjective adjustments based on period-end overall performance.

We evaluate the impact of the schemes using multiple sources of data. For tax revenue outcomes, we obtained administrative data, which we verified by conducting random spot-checks against the tax department's bank records. For outcomes such as perceived corruption and satisfaction with the tax department, we conducted a survey of over 16,000 taxpayers and their properties throughout the province. For estimating assessment accuracy, the surveyors also directly observed and recorded the property characteristics used in the tax calculation. We then manually matched surveyed properties to the tax rolls to obtain the corresponding tax records for each property. Tax assessment is determined formulaically from these property characteristics, so this allowed us to determine the accuracy of assessments by comparing our survey measurements to those on the official tax rolls.

We find that, on average across the three schemes, by the end of the two years performance pay led to an increase in tax revenue of about 9.4 log points based on the administrative data. This translates to a 46 percent higher growth rate in revenues compared to control areas. We show that this came predominantly through an increase in the reported tax base (i.e. the total assessed value of properties) rather than through increased recovery or changes in exemptions granted. On average, we find little impact of the schemes on taxpayer satisfaction. Specifically, the increased revenue generated as a result of the schemes is not accompanied by a decline in the typical taxpayers' perceptions of the quality of service from the tax office or in their satisfaction with their dealings with the tax office. We also find no overall change in the accuracy of tax assessments. Thus, on average, we find that the incentives increase revenue with little obvious downside in terms of overall perception of the tax department in the eyes of the typical taxpayer.

Comparing the three schemes, we find that they differ substantially in terms of their impact on revenue, with relatively small differences on taxpayer satisfaction and perception of the tax department. Specifically, the Revenue scheme, which provided incentives purely based on revenue collected, showed 15.2 log points higher current-year revenues relative to controls (57 percent higher growth rate) by the second year. In comparison, the Revenue Plus scheme achieved only 8.1 log points, and the Flexible Bonus scheme only a statistically insignificant 3.5 log point increase in current-year revenue. While the Revenue Plus scheme did improve perceived customer satisfaction and quality perceptions relative to to the Revenue and Flexible Bonus schemes, the differences were small, and the substantially lower revenue collected meant that this scheme had a substantially lower rate of return. The Flexible Bonus scheme did not do better on any dimension we can measure in our data, and in fact did worse compared to the control group on perception of the department's quality. Thus, adding multiple dimensions to performance pay substantially diluted the impact on revenue without a substantial corresponding increase in non-monetary outcomes.

Our survey data suggests that there was, indeed, a reallocation of rents associated with performance incentives, and finds evidence of precisely the sort of heterogeneity suggested by a simple theory of collusion. For most properties in performance pay tax circles, taxpayers were not reassessed and reported no change in tax paid. However, relative to the control group, they reported a Rs. 594 (about US $6) increase in the going rate for a bribe paid to property tax officers for properties similar to theirs, which represents a roughly 33% increase. While this does not necessarily imply that every household paid these higher bribes, respondents also indicated that bribe payments were more frequent.

However, for the small number of properties whose tax valuation was formally changed (either newly assessed or reassessed), these taxpayers report paying substantially higher taxes, but do not report the higher bribes that other properties in performance pay circles reported. Moreover, while comparisons between our survey data and corresponding administrative records suggest that typical properties are under-taxed, this does not hold for these reassessed properties, which appear on average to be taxed accurately. There is also an increase in the number of these newly assessed or reassessed properties in performance pay circles. These results are consistent with what one might expect given collusion: Performance pay means that inspectors can demand higher bribes to compensate them for their forgone (performance) pay, but, given the higher bribe now required to maintain collusion, some may instead switch from collusion (low-tax, high bribe) to non-collusion (high-tax, low bribe).

These results suggest that the increase in tax collected under the performance pay schemes is driven by a relatively small number of properties that are (correctly) reassessed and who switch from collusion to non-collusion, paying much higher taxes and lower bribes. It is interesting to examine what determines who ends up in this group. In general, we find that these newly reassessed properties have taxable value that is about 67 percent higher than the typical (non-reassessed) property. In treatment areas, the reassessed properties are even more valuable than reassessed properties elsewhere, by another 33 percent. Reassessed properties in general are also more likely to be commercial properties, which are taxed at a higher rate. There is also some suggestive evidence that, while property owners with political connections avoid being reassessed in control areas, they lose this degree of protection in treatment areas. On net, the results suggest that tax inspectors focus on a small number of high-value properties to increase revenue, thus potentially raising revenue while minimizing political costs.

From the government's perspective, the relative desirability of the schemes depends on the government's objective function. For a politician who seeks to maximize tax revenues subject to political constraints, the evidence presented here suggests that the Revenue scheme is the most effective: it raised the current-year revenue by 15.2 log points (57 percent higher growth rate), which implies a substantially positive return-on-investment (35-51%), and it did not appreciably reduce satisfaction with the tax department compared to controls. While the Revenue Plus scheme did slightly better on satisfaction than the Revenue scheme, it generated a lower (14-28%) return on investment.

This paper builds on several different literatures. First, while there is a substantial tradition of theoretical work on performance pay and compensation for tax officials in the developing world (see, for example, Besley and McLaren 1993, Mookherjee and Png 1995), there is very little empirical evidence on how these types of incentives work in practice.[1] Indeed, while there is a small-but-growing and exciting empirical literature on tax and development, it has focused to date primarily on how taxpayers respond to different types of enforcement (e.g., Gordon and Li 2009, Pomeranz 2013, Kumler et al. 2013, Carillo et al. 2014) and various aspects of the tax code (Kleven and Waseem 2013, Best et al. 2013), rather than on the role of, or how to improve performance of, tax staff. Second, this paper is related to several recent papers on improving developing country civil service performance in other contexts and using other tools. Existing work has focused on the role of wages (Dal Bó et al. 2013), intrinsic motivation (Ashraf et al. 2013), and management (Rasul and Rogger 2013). The recent work on performance pay has been centered on education and health sectors (Glewwe et al. 2010, Muralidharan and Sundararaman 2011, Gertler and Vermeersch 2013), where collusive forces are not as salient. Finally, this paper builds on the growing literature on corruption (see Olken and Pande 2012 for a review). For example, a recent paper by Duflo et al. (2013) shows that changing the incentives for third-party auditors to make them more independent increases honesty in their reporting; this paper finds similar benefits but also highlights that such incentives have the countervailing potential to increase bribes if collusion continues. More generally it underscores that when there is corruption, output-based incentives for government officials can have very different effects depending on how they affect the downstream bargaining between officials

[1] To the best of our knowledge, the best empirical evidence on the impact of performance pay on tax collection is a time-series study of a performance pay reform in Brazil (Kahn et al. 2001), which is not able to examine any non-revenue outcomes such as bribery or taxpayer satisfaction.

and citizens.

The remainder of this paper is structured as follows. Section 2 describes the relevant features of the property tax administration in Punjab, the setting in which the study takes place. Section 3 outlines theoretically what impact one might expect from performance pay in a setting with collusion between tax inspectors and taxpayers. Section 4 outlines the experimental design, Section 5 describes the data and empirical approach, and Section 6 presents the results. Section 7 concludes.

2 Setting

2.1 Property Taxes in Punjab

Punjab is Pakistan's most populous province: its population of over 80 million would rank fifteenth in the world were it a country. Property tax collection in Punjab is roughly a fifth of the level of comparable countries (World Bank 2006) due to a wide variety of problems: the tax base is narrow, tax rates do not reflect properties' market value, exemption rates are generous, tax evasion and corruption are widespread, distrust in public institutions runs high, and administration is weak (World Bank 2006, Bahl et al. 2008, World Bank 2009).

The urban property tax in Punjab is levied on the Gross Annual Rental Value (GARV) of the property, which is computed by formula. Specifically, the GARV is determined by measuring the square footage of the land and buildings on the property, and then multiplying by standardized values from a valuation table that depend only on the property location, use, and occupancy type. These valuation tables divide the province into seven categories (A to G) according to the extent of facilities and infrastructure in the area, with a different rate for each category. Rates further vary by residential, commercial or industrial status, whether the property is owner-occupied or rented, and location (i.e. on or off a main road). Taxes are paid into designated bank branches (through the National Bank of Pakistan). A copy of the receipt of payment is given to the taxpayer at the time of payment, and the bank also provides a copy to the tax collector and a copy to the provincial Treasury.

Several distortions place constraints on tax collection and introduce substantial scope for corruption. These distortions include substantially different rates for residential and commercial properties (which can be easily reclassified), as well as granting exemptions to widows, the disabled, owners of plots below 5 marlas (about 125 square meters), retired federal and provincial government employees, and religious charitable institutions (World Bank 2006). The two most notable distortions are between owner-occupied and rented residential properties (the latter are taxed ten times more) and between residential and commercial properties (the latter are taxed between 3 and 6 times more). Qualitative evidence suggests that these distortions are the main ways in which tax evasion takes place, both due to the significant impact these margins have on tax assessment and also because it is less easy to verify whether a residential property is being rented or, particularly for mixed usage properties, what fraction of the property is being used for commercial purposes.

For research purposes, a methodological advantage of property taxes is that, unlike most taxes, true property tax liability can be independently estimated by the researcher. By comparing official tax payments to an independent assessment by an external survey team, we can determine changes to both the accuracy of tax evasion and the average level of over or under taxation. This approach follows other examples in the corruption literature (e.g., Fisman and Wei 2004, Olken 2007).

2.2 Property tax administration

The primary unit of tax collection is the "tax circle," a predefined geographical area that covers anywhere from two to ten thousand unique properties. Within each circle there are three designated tax officers who work together as a team: an "inspector" who leads the team and determines tax assessments and issues notices that demand payment; a "clerk" who is in charge of record keeping; and a "constable" who assists the inspector in the field.[2] Together they maintain a record of all properties and their attributes (size, type of use, etc.), apply the valuation tables to each property, and determine which property tax rates and exemptions apply to the property. Following this process, the inspector determines each property's tax liability and sends an annual demand notice to the property owner for payment at a bank.

All three officials are part of the provincial career bureaucracy, with wages determined by salary band and length of service. As is common for civil servants in developing economies, tax officials receive fairly low wages that are rarely, if ever, tied to performance. However, since the department has explicit financial targets each year, there is pressure on each circle team to contribute. This occurs typically through each administrative level pressuring lower levels to increase collections. With limited reward mechanisms and vertical mobility, threats of transfers are the primarily tool available to supervisors who want to improve performance. While some inspectors do have strong preferences over their posted circle, these threats have limited effectiveness since transfers are often more politically based than merit based (Piracha and Moore 2015).

The problems from the lack of an explicit and transparent performance based reward system are exacerbated by the fact that the system leaves considerable opportunities for leakages, collusion, and low collection, especially because there are few independent checks on the actions of the tax circle team and limited audit mechanisms. The property database is manually recorded on physical registers and does not automatically include new properties or property updates. Building permits and rental agreements are not always formally registered, and when they are registered they are not automatically linked to tax rolls, so there is no way for the tax department to learn about new construction or changes in property use except through the efforts of the circle staff. In addition, officials may employ significant discretion in applying valuation tables to individual properties and

[2] In practice, while a circle will almost always have an inspector, at times the clerk or constable position may remain vacant due to hiring frictions and freezes. At baseline, 46% of circles were missing either a constable or a clerk.

determining exemptions. For example, properties can be incorrectly designated as owner-occupied when they are being rented out (and as noted above, the latter are taxed at a ten times higher rate), classified as residential when they are in fact commercial, designated as "off road" when they are on a main road, or mis-measured. Finally, the manual system of billing and collection, in which tax bills are hand-written by inspectors and clerks and hand-delivered by tax constables, is prone to errors and/or manipulation in crediting collections. Given the incompleteness of property records, the complexities of joint family property rights, the informality of most rental arrangements, and the fact that the manual official records are only kept at the circle level, it is extremely difficult to verify how the circle level tax officials use their discretion.

In this context, performance pay has the potential to induce tax officials to raise collections. While this could be due to greater effort in tracking new properties and uncovering physical and usage changes that increase tax assessments, anecdotal evidence suggests that collectors likely have substantial private information regarding a property's true tax liability already, which they use for extracting bribes rather than assessing higher taxes. They may choose to only reveal (parts of) this information to the authorities when faced with significant opportunity for rewards. An extreme form of such information disclosure is revealing the existence of (newly constructed) properties and formally adding them to tax registers (recall there is no automatic process though which this happens). In addition, tax collectors may increase valuations by revealing the true, higher tax valuation of a property or denying (incorrectly provided) exemptions. The next section formalizes these incentives to strategically disclose information within a standard model of collusion.

3 Theory

Consider a simple setting where a taxpayer i faces a true tax liability τ_i^*. The tax inspector knows τ_i^*, but can choose to report a lower tax liability to the government, τ_i, instead if he so chooses. The tax inspector receives an incentive payment which is a constant fraction r of actual taxes paid, i.e. $r\tau_i$.

Both taxpayer and tax inspector face costs from colluding to report $\tau_i < \tau_i^*$. The taxpayer's cost of accepting a reduced tax liability is $\alpha_i (\tau_i^* - \tau_i)$ and the tax inspector's cost of giving a reduced tax liability is $\beta_i (\tau_i^* - \tau_i)$. [3] We allow α_i and β_i to differ for each individual taxpayer-inspector pairing i.

We assume that the taxpayer and the tax inspector engage in Nash bargaining, with the taxpayer potentially paying a bribe b_i as a transfer to tax inspector. If no agreement is reached, the taxpayer

[3] It is also worth noting that while the costs are modeled in terms of deviations from the true tax liability, i.e. $\alpha (\tau_i^* - \tau_i)$, an alternative formulation would be to have costs in terms of bribes paid and received, i.e. to value bribes $\alpha_i b_i$ instead of b_i, so that bribes are less valuable than cash. This could represent the fact that there is some chance bribes are detected, or that one needs to launder bribe money to evade detection, which makes it less valuable than legal money. In Appendix A.2, we show that we obtain the same qualitative results as in the model derived here if we specify the costs in terms of bribes rather than in terms of tax evasion.

receives payoff $-\tau_i^*$ and the tax inspector receives payoff $r\tau_i^*$. If an agreement is reached, the taxpayer receives payoff $-\tau_i - \alpha_i (\tau_i^* - \tau_i) - b$ and the tax inspector receives payoff $r\tau_i - \beta_i (\tau_i^* - \tau_i) + b$.

To arrive at the solution, note that the joint surplus from agreement is

$$\tau_i^* - \tau_i - \alpha_i (\tau_i^* - \tau_i) + r(\tau_i - \tau_i^*) - \beta_i (\tau_i^* - \tau_i) \tag{3.1}$$

which can be rewritten as

$$-\tau_i (1 - \alpha_i - \beta_i - r) + (1 - \alpha_i - \beta_i - r)\,\tau_i^* \tag{3.2}$$

This equation shows that if

$$r + \alpha_i + \beta_i < 1 \tag{3.3}$$

the joint surplus is maximized at $\tau_i = 0$ (full collusion); otherwise the joint surplus is maximized at $\tau_i = \tau_i^*$.

Suppose that γ_i is the bargaining weight of the taxpayer (and $1 - \gamma_i$ is the bargaining weight of the inspector). If collusion takes place, the bribe paid is such that each side receives their outside option plus their share of the surplus. This implies that the bribe the taxpayer pays to the tax inspector is

$$b_i = \left[(\beta_i + r)\,\gamma_i + (1 - \gamma_i)(1 - \alpha_i)\right]\tau_i^* \tag{3.4}$$

What are the implications for tax revenue and bribes of moving from no incentive ($r = 0$) to positive incentive payments r? This simple framework shows that it depends on whether the equilibrium shifts from the collusive equilibrium to the non-collusive equilibrium. So long as $r + \alpha_i + \beta_i < 1$ and $\gamma > 0$, increasing the incentive rate increases bribes (since the taxpayer now has to compensate the inspector for the foregone incentive payments). On the other hand, if increasing r means that the threshold is crossed such that $r + \alpha_i + \beta_i > 1$, then collusion disappears, bribes fall to zero, and tax revenue increases from 0 to τ_i^*. The result that bribes increase with incentives to the tax collector, but that collusion may disappear if incentives are sufficiently great, is closely related to Shleifer and Vishny (1994) and Boycko et al. (1996), who study bribes between politicians and managers of firms as part of their analysis of privatizations.

Since there is heterogeneity in α_i and β_i, the aggregate response of tax revenue will depend on the fraction of households induced to switch from the collusive to non-collusive equilibrium. Denote by $f(\alpha, \beta, \tau^*)$ the joint distribution of α, β and τ^* in the population. Then the increase in total tax revenue, T, in response to an increase in r is given by

$$\frac{dT}{dr} = \iiint\limits_{r+\alpha+\beta=1,\tau^*\in(0,\infty)} \tau^* f(\alpha, \beta, \tau^*)\, d\alpha\, d\beta\, d\tau^* \tag{3.5}$$

Equation (3.5) makes clear that the change in tax revenue depends on the density of taxpayers that are just indifferent between colluding and not colluding, and the average tax liability of those households.

The welfare implications depend on the degree to which the costs from avoidance, α and β, are social costs or private costs (Chetty 2009). To the extent that they represent social costs (i.e. they are utility costs from being dishonest or caught cheating, and so represent real costs, as opposed to fines, which would just be transfers), then assuming equal welfare weights on taxpayers and tax inspectors, the social welfare gain from the increase in r is equal to the increase in tax revenue less the cost of the incentive payments.[4] To see this, write social welfare as the net surplus of the taxpayer, tax inspector, and government for each taxpayer-tax inspector pair i, as follows:

$$W = \int_i \left[\underbrace{-\tau_i - \alpha_i \left(\tau_i^* - \tau_i\right) - b}_{taxpayer} + \underbrace{r\tau_i - \beta_i \left(\tau_i^* - \tau_i\right) + b}_{taxinspector} + \underbrace{\tau_i - r\tau_i}_{government} \right] f(\alpha, \beta, \tau^*) \tag{3.6}$$

Assuming equal social welfare weights, when we change r, the only change in social welfare comes from those taxpayer-tax inspector pairs i induced to switch from collusive to non-collusive; for pairs that remain collusive, bribes increase, but this is just a transfer, and does not affect social welfare. Similarly, for pairs that remain in the non-collusive equilibrium, there is no change in overall welfare since overall taxes paid are unchanged (and there are no bribes) and the greater incentive payments made to the tax collector are entirely offset by the cost of these payments to the government. Those taxpayer/tax inspector pairs induced to change by a marginal increase in r are those that were just indifferent between colluding and not-colluding (i.e. had $\alpha_i + \beta_i + r = 1$), so a switch from collusion to non-collusion does not change the sum of taxpayer and tax inspector utility. The government, however, experiences a first-order utility change equal to the tax revenues it collects, less the incentive payment it needs to pay out. The fact that social welfare is equal to the net change in the government's fiscal position is related to the classic result by Feldstein (1999), with the exception that in this case, the "fiscal externality" is tax revenue net of incentive payments.

The model presented here was simplified for ease of exposition, in that the costs to reducing tax liability are linear. Linearity is not crucial for the main qualitative results; as we outline in a more general model (which also avoids the corner solutions inherent in the linear case) in Appendix A.1, all we need for the qualitative patterns we discuss is that the marginal costs of collusion to both parties are weakly positively increasing in $\tau_i^* - \tau_i$. [5] The key difference in this generalized model is

[4] For example, in Pakistan, if one is caught cheating on taxes, the primary action taken is that a property might be sealed so no one could use it until the back taxes are paid; this represents a real social cost, since the property is unusable during this period. If the government has different welfare weights for payments from taxpayers, bribes received by taxpayers, and tax revenues received by the government, then the welfare formula would be more complex and would need to take these differences into account.

[5] In the linear model, conditional on colluding, one always sets $\tau_i = 0$, so the only reason taxes increase in this

that the increase in revenue from an increase in r comes not just from households that switch from collusive to non-collusive (as in equation (3.5)), but also from households that continue to collude, but now collude a bit less than before. Welfare analysis, however, will be similar.

Note that by assuming that the tax inspector knows τ^*, we have suppressed both an effort and an overtaxation/extortion margin. The effort margin recognizes that with more effort tax inspectors could discover more properties or learn a property's true tax liability. We suppress it in the model since models of increased effort under incentives are well understood, and we wish to focus on the bargaining implications, though changes in effort could be possible in our context as well. We give one example in Appendix A.2 of how our framework could be extended to include an effort component. As we show in that case, including an effort margin does not yield any qualitatively different insights. While the overtaxation margin is conceptually interesting, in practice this appears less common (as we will show below, our property survey suggests the typical property is in fact under-taxed, not overtaxed) and we therefore do not incorporate it in the model.

4 Design

This section presents the design of the performance pay mechanisms introduced and the experimental design of the study. Section 4.1 describes the performance pay program, and Section 4.2 describes the randomization and balance check.

4.1 Performance pay design

Tax circles were randomly allocated into one of three performance pay schemes: the Revenue, Revenue Plus, and Flexible Bonus schemes. A total of approximately 70 circles were allocated to each of these three schemes (50 each in the first year and an additional 20 each in the second year). In addition, in the second year of the experiment, two new treatments were added: a performance pay scheme for supervisory personnel, and an "information-only" scheme that replicated the information, meetings, and perceived salience of the Revenue scheme, but without any financial payments. We describe each scheme below and then conclude with a brief discussion of how well tax officials understood the schemes and the degree to which they believed the schemes would be implemented as described.

4.1.1 Revenue-based

This performance pay group rewarded tax circle staff (inspectors, constables, and clerks) based on the revenue they collected above a predefined benchmark. The benchmark for each circle was

model is a shift from full collusion to no collusion as r increases. In a more general model, there could also be interior solutions for τ_i, and so τ_i could also change in response to r. In Appendix A.1 we allow for more general cost functions. We show that in this more general case, the results shown here continue to apply: as r increases, tax revenue continues to increase, but the impact on bribe payments is ambiguous.

generated using historical revenue data for that circle. Specifically, each inspector continued to receive his or her current base salary, plus a bonus calculated by the following formula:

$$Bonus_c = \alpha_c \max(Revenue_c - Benchmark_c, 0) \tag{4.1}$$

where the bonus rate α_c is 40% for those circles below the 50th percentile in baseline revenue, 30% for those circles between the 50th and 75th percentiles in baseline revenue, and 20% for those circles above the 75th percentile in baseline revenue. The differential bonus rates were put in place for equity considerations, i.e. staff in larger circles were compensated at a lower rate than those in smaller circles, where it was perceived to be more difficult to raise a given amount of revenue. It is important to note that this scheme treated increased collections due to expansion of the tax base (new properties) or increased collection on the current base (higher recovery rates) symmetrically. Benchmarks were generated using a three-year average of historical collections, adjusted for the normal rate of increase in collections, and were designed such that most circles would be "in-the-money" and face linear incentives on the margin.[6] Since most inspectors are rotated to new circles every two to three years, the use of 2 to 4 lags of revenue collection in determining benchmarks means that ratchet effects should not be a first-order concern in this context. This is because by the time higher revenue collection starts to impact benchmarks substantially, the inspector would likely be in a different circle and not subject to those benchmarks.

As each tax circle staff consists of three members, the bonus was divided 40%-30%-30% among inspector, constable, and clerk, respectively. On net, with a 30% average incentive payment to the group, and this division among the three group members, each individual inspector, constable, and clerk faced a roughly 10% individual marginal incentive. Payments for all incentive schemes were restricted to staff who were posted in the circle at the time of randomization, and staff were no longer eligible to receive payments if they were transferred to a non-incentivized circle.

4.1.2 Revenue Plus

The Revenue Plus scheme was similar to the Revenue-based scheme, but included additional incentives (the "plus" component) to help address the multitasking problem inherent in the tax

[6] Specifically, in the first year (FY11-12), the historical benchmark was the three year average of revenues from FY07-08, FY08-09, and FY09-10, plus 10%. Since the rate of increase in collections averaged about 8% per year, the benchmark should be approximately 13% below the average revenue under business-as-usual. This was done by design so that almost all circles (even those with lower than average collections) would be in-the-money and face linear incentives on the margin (Holmstrom and Milgrom 1987). The adjustment rate was increased slightly in Year 2 in light of the growth rates observed in Year 1, so that in the second year (FY12-13), the historical benchmark was the three year average of revenues from FY08-09, FY09-10, and FY10-11, plus 20%. We should also note that in the first year of incentives, there were separate benchmarks for current-year tax collection and arrears collection, so that the formula was $Incentive_c = \alpha_c \max(CurrentYearRevenue_c - CurrentYearBenchmark_c, 0) + \alpha_c \max(ArrearsRevenue_c - ArrearsBenchmark_c, 0)$. Given that inspectors have some leeway in classifying revenue into current or arrears, but no flexibility in total revenue (since it must match the amount of money deposited into the bank), in the second year, incentives were simplified to be based simply on the total revenue collected.

collector's job (Holmstrom and Milgrom 1991). Specifically, in addition to maximizing revenue collected, the government also cares about the costs of taxes in terms of how people feel they are being treated by the tax department and whether taxes are being assessed accurately.

To address these concerns, in addition to rewarding on revenue using the exact same formula as in the Revenue scheme, this scheme adjusted pay based on taxpayer satisfaction and accuracy of tax assessments. Circles in the scheme were ranked based on the accuracy and satisfaction measures and divided into three equal-sized groups. Circle staff were paid as in revenue treatment, but the top group received an additional bonus equal to 0.75 times their base salary, and the bottom group lost 0.75 times their average base salary.[7] By design the total payments under the scheme could never be negative (that is, base salary was never at risk; an inspector in the bottom group might receive 0 from the scheme but would never forfeit his base salary); otherwise, (conditional on the same revenue increase) average payments would be identical between the Revenue and Revenue Plus schemes.[8]

The satisfaction and assessment accuracy measures were based on an independent survey of 12,000 randomly sampled properties (described in Section 5.1 below). Taxpayer satisfaction was measured based on two survey questions about the quality and results of interactions with the tax department.[9] Accuracy was measured as 1 minus the absolute value of the difference between GARV as measured by the survey and the official GARV, as measured from the tax department's administrative records, divided by the average of these two values. Since this "Plus" component relies on third party surveys and could also lead to losing the performance pay earned due to increased tax collections, it effectively constitutes an audit component (though was not referred to as such so as to maintain better optics).[10]

4.1.3 Flexible Bonus

The third scheme was designed to be analogous to the way bonuses work in the private sector for many complex jobs, such as those in Wall Street firms: managers distributed a fixed bonus pool to

[7] Inspectors in the top group received an extra Rs. 15,000 per month, and constables and clerks received an extra Rs. 11,500 per month; those in the bottom group lost an equivalent amount.

[8] To the extent that tax officials are risk-neutral, or they are risk averse but have CARA utility with cost of effort expressible in monetary terms, the additional variance induced by the Plus scheme should not affect their choices other than through the multi-tasking channel. Relaxing the assumptions of risk-neutrality or CARA utility, however, could allow there to be direct effects from the increased variance due to the Plus component on the return to revenue from the effort component.

[9] The questions were "In your opinion, what has been the overall quality of service offered by this department to this property?" and "In your personal dealings with members of this department, how satisfied are you with the outcomes?" Each question was answered on a 1 to 5 Likert scale.

[10] In the first year, this measure was noisier due to survey and measurement logistics that were resolved by the second year. Therefore in the first year we instead calculated accuracy by correlating Log GARV in the official register with Log GARV according to the survey, which was more robust to being off by a constant. The inspector's accuracy sub-score was based on the strength of this correlation.

talented employees based on all factors (including subjective ones) they observe.[11] In this treatment, staff were again divided into three groups and pay was determined by group (just as in the Revenue Plus scheme), but rather than have their pay determined by an ex-ante specified formula, they were divided by their performance as ranked by a departmental "Performance Evaluation Committee" (PEC) comprised of senior tax officials. Everyone in the treatment provisionally earned a base salary supplement roughly equal to their average salary.[12] At the end of the year, adjustments were made just as in the Revenue Plus scheme: the top third of circles received an additional bonus equal to 0.75 times their base salary, and the bottom group lost 0.75 times their average base salary.[13]

In determining payments under this scheme, the PEC was allowed to use any criteria it chose, so long as it could document a reason behind them, and the committee was provided all of the same information used in the Revenue Plus treatment (increase in revenue over benchmarks, customer satisfaction, and accuracy of assessments). The main differences between the Flexible Bonus and Revenue Plus schemes were that the objective revenue-based formula was replaced by a fixed increase in base salary (with an end of year bonus), and that the grouping was made by the Performance Evaluation Committee as it saw fit with few restrictions, rather than being a mechanical formula based on customer satisfaction and accuracy.

Although the official design of the treatment allowed the PEC full flexibility in using subjective criteria, they in fact created a (richer) formula for ranking circles, using the following indicators and weights (in parentheses): increase in revenue collected (40 percent), increase in tax base (25 percent), accuracy of assessment (15 percent), subjective director's rating (10 percent), and customer satisfaction (10 percent). This was publicized about 6 months after the intervention began, so by the beginning of Year 2, inspectors should have been fully aware of the assessment criteria. The two additional criteria (compared to Revenue Plus) included tax base increases and the subjective director's assessment. On net, the correlation between the Performance Evaluation Committee ranking and the ranking of payments that would have been generated under the Revenue Plus

[11]For example, managers might be able to observe effort in addition to outcomes; they also might have information that certain areas were more difficult than others, and so could adjust for these factors in ways that would be difficult in an objective, ex-ante specified, formulaic incentive system. While such subjective assessments can potentially better match the complexities of real jobs, they can be less effective than formulaic systems if workers do not trust the managers to implement them properly, if managers play favorites, or if managers and workers disagree about the subjective component of performance (Baker et al. 1994, Prendergast and Topel 1996, Prendergast 1999, MacLeod 2003).

[12]In the first year of the project, the base salary supplement was Rs. 30,000 for inspectors and Rs. 23,000 for constables and clerks. This amount was closer to one and a half times base pay. However, this figure was adjusted in the second year to Rs. 22,000 and Rs. 16,500 in order to better ensure that the three schemes generated equal average honorariums.

[13]Inspectors in the top group received an extra Rs. 15,000 per month, and constables and clerks received an extra Rs. 11,500 per month; those in the bottom group lost an equivalent amount. In practice, since it was not feasible to actually take money back once paid, circle staff were only paid 50% of the honorarium earned each quarter (this was also true of the two other schemes) and then adjustments were made at the end of the year (in Year 1) and at both the half-year and year end points in year 2; see Section 4.1.5 below.

formula was 0.269 in Year 2.

4.1.4 Additional treatments

In addition to the main circle-level performance pay treatments, which were in place for two fiscal years, we introduced two additional treatments in the second year of the program (FY12-13). The "information-only treatment" was intended to capture the part of the effect that arises from all other aspects of treatment besides the monetary incentives. Seventy circles were randomized into this treatment. Staff from these circles went through the same process as the staff in the Revenue treatment (including receiving quarterly reports on their collections above their historically-predicted benchmarks and attending quarterly meetings to review their progress), but with no corresponding incentive payments. While the quarterly reports just repackaged information that staff already had, the reports presented the information in a more systematic format, which may have increased its salience. Furthermore, the act of attending the quarterly meetings may have led circle staff to believe that they were being monitored more carefully. The information-only scheme therefore nets out these effects from the direct impact of the payments per se in the performance-based incentives.

In addition, a supervisor's performance pay scheme was introduced in the second year. This was identical to the Revenue scheme, except that it applied to both the Assistant Excise and Taxation officers (AETOs), who supervise the circle staff, and the Excise and Taxation officers (ETOs), who supervise the AETOs. Randomization was done at the level of the ETO, with 26 treatments and 25 controls. All AETOs working under selected ETOs were included. Payments were calculated based on the average increase in revenue over benchmarks for circles under their supervision. The bonus rate was determined by average circle size, and each supervisor received a 50 percent share of all imputed bonus payments (recall an inspector's share was 40%). Since this intervention was randomized at the level of the ETO (of which there are only 51 in the province), whereas the circle-level intervention was randomized at the circle level (with almost 500 circles), this intervention will have substantially lower statistical power than the main circle-level treatments.

4.1.5 Knowledge and Credibility

In order to ensure that collectors understood the specifics of the scheme they were in, we carried out detailed trainings for each scheme at the start of the year, followed by post-training quizzes and refresher trainings throughout. By seven months after treatments started, quiz results revealed that virtually all inspectors were able to understand the scheme and accurately calculate the payments to which they would be entitled. An independent survey of all inspectors (treatment and control) confirmed that inspectors could accurately identify whether they would receive payments, and which scheme they were in. To ensure that inspectors believed that payments would actually be made, the project was officially approved by the Chief Minister (the highest political authority in the province). A small pilot was conducted (and payments made) in 11 circles for an entire year before the main experiment began, and payments were made quarterly throughout the main experiment.

4.2 Randomization Design and Balance Checks

The randomization was carried out through public lotteries, with representatives from the tax department present. This helped minimize any perceived bias, especially since the performance pay schemes were popular (most staff wanted to opt-in). In order to reduce any concerns about differential selection across the schemes while maintaining informed consent, the lottery was conducted in two stages. In the first stage, circles were selected to participate in the project and staff consent to participate was sought. Staff were told about the three possible incentive schemes, and it was made clear that a second lottery would determine which scheme they would be assigned to.[14] Once consent was obtained, a second lottery was held to assign consented circles into particular incentive schemes. Over 95% of circle staff that were selected in the first lottery consented to participate. Given the extremely high consent rates observed in the first year, both stages were conducted in a single lottery in Year 2. The lotteries were held as close as possible to the start of the fiscal year on July 1.[15]

Table 1 shows the experimental design. In Year 1 of the program, a total of 160 circles were selected in the first ballot, to be divided equally into one of three treatments. In Year 2 of the program, an additional 58 were selected and divided into the same three treatments. The circles selected in Year 1 remained in their same treatment assignments, and new inspectors who had previously transferred into these circles became eligible for performance pay in Year 2.[16] In addition, 70 circles were selected for the information-only treatment. Each of these lotteries was stratified with 19 strata based on the 11 administrative divisions of the province and – for all but the smallest few divisions– circle size.

Appendix Table 11 compares the selected circles to controls on their baseline characteristics in the administrative data (described in Section 5.1 below) based on the final randomization at the end of Year 2. Out of the 42 comparisons made (7 variables * 6 columns), only 1 is significant at the 5 percent level (the coefficient on log non-exemption rate for the Flexible Bonus scheme).[17]

[14] Given the crucial role played by the inspector in collecting tax, it was decided that the circle as a whole could only participate if the inspector consented to participate. Constables or clerks could individually opt out of the scheme as they saw fit. This, however, rarely happened.

[15] In the first year, the 1st stage lottery was held on July 9, 2011; after consent was obtained, the 2nd stage lottery was held on August 10, 2011. In the second year, the lottery was held on July 7, 2012.

[16] Since this was not part of the policy initially (we had made clear that anyone transferring in during the year would not be part of the treatment) there is not much concern that staff were strategically transferring in the hope that they would be eligible in the second year.

[17] Looking scheme by scheme, the joint test for statistical balance shows statistical significance in one of the schemes (Revenue Plus) compared to pure controls, even though none of the individual covariates are statistically significantly different. In the Appendix Tables we show that the main average effects of incentives do not seem to be driven by

5 Data and Empirical Methodology

5.1 Data

We use two main sources of data for analysis: circle-level administrative data for our main measures of tax performance, and property/taxpayer-level data based on a survey we conducted to obtain measures of accuracy of tax assessment, customer satisfaction, and corruption. Appendix B provides further details on both datasets. In particular, it outlines the additional verification and checks we ran on the administrative data, including how we addressed complications that arise for tax circles that experience boundary changes over time, and notes the details of the survey exercise including variable construction for key outcomes of interest. Here we simply highlight a few of these aspects.

The administrative data is based on the quarterly reports that each inspector files, which show their overall collections (separately for current year and past years/arrears collections) and the total assessed tax base. As detailed in Appendix B we digitized these reports for all tax circles and selected a random sample to be verified in each of our project years. The verification was done by aggregating (thousands of) bank-verified receipts of individual payments in a given tax circle. We found no statistically or economically significant discrepancy between the administrative data and our independent verifications.

Summary statistics for key variables from the administrative data are shown in Panel A of Table 2 for the second year of the experiment (FY 2013); summary statistics for additional years and variables can be found in the Appendix. Several observations are worth noting. First, current year revenues are substantially larger than arrears (i.e. collections against past years' unpaid taxes) – the mean of log current revenues is 15.52 compared with just 13.91 for log arrears, implying that, on average, current revenue in the typical circle is about 5 times as large as arrears. This suggests that the main impacts on total revenue will likely be felt through increases in current year revenue. Second, there is much more variation in arrears – the standard deviation in log arrears is about 1.5 times that of log current revenue – implying that detecting effects on arrears statistically will be more difficult. It is also interesting to note that the log recovery rate (the log of tax revenue divided by the tax base net of exemptions) is -0.14 for current year taxes, which this implies that about 85 percent of all taxes that are demanded by the government are in fact paid. Thus while non-payment is a substantial issue (a typical developed country government would not be satisfied with a 15 percent non-payment rate of property taxes),[18] it is still the case that the bulk of taxpayers do

this one sub-treatment (Appendix J), and that controlling for the variables included in the balance table does not meaningfully change the results (Appendix K). The Appendices can be found at `http://goo.gl/1xifaf`.

[18] For example, on average from 2010 to 2013, the city of Cambridge, Massachusetts collected almost exactly 100 percent of all property taxes due. Even excluding reductions in taxes due to abatements (e.g. for poor households), it collected 98.5 percent of the pre-abatement gross property tax levy (City of Cambridge 2014). The collection rate in Pakistan is more comparable to Detroit, Michigan, which had an 80 percent average property tax collection rate from 2010 to 2013 and is currently legally bankrupt (City of Detroit, Michigan 2013).

in fact pay the tax bills they receive. Thus any potential evasion may come from under-assessment of properties (as we will see below) rather than just flagrant disregard of issued tax notices.

The second primary data source is the property survey we conducted at the end of the two year period. This survey provides our main non-revenue outcomes (taxpayer satisfaction measures and tax assessment accuracy), as well as owner/property characteristics that help us examine any heterogeneous effects. The survey is based on two distinct samples. The first, which we will refer to as the "general population sample," consists of roughly 12,000 properties selected by randomly sampling 5 GPS coordinates in each circle and then surveying a total of 5 (randomly chosen) properties around that coordinate. These properties therefore represent the picture for the typical property in a tax circle. The second sample, which we will refer to as the "reassessed sample," consists of slightly more that 4,000 properties (roughly 10 per circle) which were sampled from an administrative list of properties that are newly assessed or reassessed. These properties were then located in the field and surveyed. The purpose of this survey was to over-sample the (few) properties that experience such changes each year so as to be able to examine the impacts on such properties separately (see more on this in the next empirical methodology below).

Panel B of Table 2 presents summary statistics for properties from the general population sample. Several facts are worth noting. First, observe that 84 percent of properties we randomly sampled in the field were successfully located on the tax registers. Again, while there are a substantial number of untaxed properties, it is not the case that only a few properties are on the tax rolls. Second, conditional on being on the tax rolls, on average properties appear under-taxed. We focus on the Gross Annual Rental Value ($GARV$) of the property, which is the main measure of a property's tax value, before exemptions and reductions are applied.[19] To measure under or over taxation, we focus on the "tax gap," defined as

$$TaxGap = \frac{GARV_{Inspector} - GARV_{Survey}}{(GARV_{Inspector} + GARV_{Survey})} \tag{5.1}$$

This captures the difference between what the inspector officially reported and what was obtained through our own survey. Our measure of inaccuracy is the absolute value of the tax gap (for more details see Appendix B). On average, inaccuracy is 0.34, indicating substantial disagreement between the two measures. The tax gap has a mean value of -0.10, suggesting that under-taxation is prevalent in our population.[20]

Corruption also appears to be prevalent. On average, respondents report that annual bribes paid

[19]We focus on GARV, rather than tax assessed, because nonlinearities in the tax formula mean that there is substantially more measurement error in tax assessed than in GARV. For example, if the land area is less than 5 marla (1,361 square feet), non-rented, residential properties are completely exempt from tax. By contrast, GARV is a continuous function of the underlying property characteristics and hence is much more robust to measurement error.

[20]Given the way it is normalized, an average Tax Gap of -0.10 means that on average the inspector's assessment is 19% less than the survey's estimate. See Appendix B for more on this.

for a property similar to theirs are around Rs. 2,000 (US $20) – about half of the amount they report paying in property taxes. Bribes are frequent – when asked how many times a typical property owner would need to bribe the property tax department, the mean is 0.76 bribes paid per year. On the other hand, respondents are not wildly unsatisfied with service from the tax department – on a 0-1 scale, the average response is 0.53 for quality of service and 0.55 for satisfaction.[21] Of course, this could be consistent with corruption: a respondent might be "satisfied" if he was able to reduce his official tax liability by paying a bribe.

In addition to these two primary sources of data, in some of the appendix tables we also make use of a short phone-based survey of inspectors where we gathered basic information about the self-reported effort and perceived supervisory support and pressure felt by the tax inspectors.

5.2 Empirical Methodology

Since we are evaluating a randomized experiment, the empirical methodology is straightforward. We estimate 2SLS regressions, where the endogenous variable is the treatment status at any point in time and the instruments are the results of the lottery.[22] Specifically, our primary specification for assessing circle-level outcomes using the administrative data is

$$\ln Y_{cst} = \alpha_s + \beta Treatment_{cst} + \gamma \ln Y_{cs0} + \epsilon_{cst} \tag{5.2}$$

where Y_{cst} is the outcome of interest for circle c in stratum s at time t, and $Treatment_{cst}$ is a continuous variable that takes values from 0 to 1 that represents the fraction of treated circle staff present in circle c in the last quarter of the given fiscal year. Y_{cs0} is the value of the outcome variable at baseline (i.e. in the fiscal year prior to randomization). $Treatment$ is instrumented by a binary variable that represents the circle's randomization status into any one of the three incentive schemes.[23] We include stratum fixed effects (α_s) given the lottery was stratified by these strata. All regressions based on administrative data are run using circle boundaries that existed at the time of randomization. We report robust standard errors clustered at the level of the robust partition of circles, i.e. the maximum set of circles that have been involved together in a set of

[21] One might be concerned that the quality and satisfaction variables are simply picking up noise. However, Panel A of Appendix Table 12 shows that the satisfaction and quality measures are internally consistent: that is, households who report higher satisfaction report higher quality of service, and households that report higher quality report lower bribes, and so on. More importantly, households in a circle tend to agree with each other. Panel B of Appendix Table 12 regresses these measures on what other respondents in the same circle report: people report high satisfaction when others in their neighborhood report high satisfaction, report high bribes when others report high bribes, and so on.

[22] The reason the treatment status is not exactly equal to the lottery results is that a small number of circles (8 out of 482) did not consent to participate, and because some circle staff lost eligibility to continue in the scheme after they were transferred out to another circle.

[23] Reduced-form versions of the main table can be found as Appendix Tables 3-G1 and 4-G1. Note also that the information-only scheme is not included as a treatment, but is instead included as part of the control group to maximize statistical power. Appendix Tables 3-I through 7-I re-estimate the tables in the paper where, instead, the information treatment is separated out, so performance pay treatments are compared only to pure controls, with qualitatively similar results.

splits and merges since randomization.

To estimate the impact of the separate sub-treatments, we estimate the analogous regression separately by treatment:[24]

$$\ln Y_{cst} = \alpha_s + \beta_1 Revenue_{cst} + \beta_2 RevenuePlus_{cst} + \beta_3 FlexibleBonus_{cst} + \gamma \ln Y_{cs0} + \epsilon_{cst} \quad (5.3)$$

For survey-based outcomes, we run regressions at the individual property level. As discussed above, we have two separate samples, the general population sampled from random GPS points, and properties that were sampled because they had a change in their tax assessment (either previously assessed properties that were reassessed, or properties newly added to the tax rolls). When examining the general population sample, we run regressions of the form:

$$Y_{ics} = \alpha_s + \beta Treatment_{cs} + \epsilon_{ics} \quad (5.4)$$

where i is an individual property. As above, we instrument for $Treatment$ with the randomization results.[25] We include stratum fixed effects and cluster standard errors at the circle level. When available, we include controls for baseline level outcome variables.[26]

For regressions where we are interested in the difference between reassessed and new properties and regular properties, we include both samples, and then run regressions of:

$$Y_{ics} = \alpha_c + \beta_1 Treatment_c * ReAssessed_{ic} + \beta_2 ReAssessed_{ic} + \epsilon_{ic} \quad (5.5)$$

where $ReAssess$ is a dummy that is 1 if a property was sampled from the list of properties whose valuation was changed (we do not distinguish in this regression between properties whose tax valuation was changed and newly assessed properties; both are captured by $ReAssessed$). Note

[24]In such regressions, in addition to reporting β_1, β_2, and β_3, we report several other statistics to guide the analysis. In particular, we report the p-values for a test of the joint statistical significance of the incentive schemes (i.e. a test of the null that $\beta_1 = \beta_2 = \beta_3 = 0$) and a test that the three schemes are identical (i.e. a test of the null that $\beta_1 = \beta_2 = \beta_3$). We also report p-values from a test of whether the schemes that dealt with multi-tasking are identical to those that did not (i.e. a test of the null that $\beta_1 = \frac{\beta_2+\beta_3}{2}$), and from a test of whether the scheme that used subjective information from the department is identical to the formulaic schemes (i.e. a test of the null that $\beta_3 = \frac{\beta_1+\beta_2}{2}$).

[25]Regressions based on survey data are run using circles boundaries when the sample of properties was drawn, which happened in the middle of the second fiscal year of the study.

[26]As discussed above, our sampling strategy was to randomly draw 5 initial GPS coordinates from within the boundary of a tax circle. We then survey the property closest to that point and then following a left-hand rule (or if that is not possible, a right hand one) survey an additional four properties. A potential concern is that we may be oversampling larger properties since a randomly chosen GPS point is more likely to fall inside a larger property. While this may be true for the first sampled point, we have confirmed that it is not true of subsequent properties i.e. there is very little correlation between the land area of the first property (chosen by GPS point) and the subsequent properties (chosen by moving to the left). As a robustness exercise we therefore redo our estimates after dropping the first sampled point and using only the remaining points, and find that our results are qualitatively similar. See Appendix Tables 8-L and 6-L.

that unlike equation (5.4), we now include circle fixed effects (α_c) to capture fixed differences among circles between properties. We examine the analogue of equation (5.3) when we examine sub-treatments.

In interpreting equation (5.5), it is important to note that which properties are reassessed is, of course, potentially an outcome of the treatment as well. As such, the coefficient β_1 includes two margins of treatment effects – an extensive margin effect (i.e. the type/number of properties revalued can be impacted) and an intensive margin effect (a given reassessed property may now be dealt with differently). For example, if Y_{ics} is the amount of bribes paid, the coefficient β_1 in equation (5.5) shows how the difference in bribes paid between reassessed and non-reassessed properties changes in treatment versus control circles. As outlined in the conceptual framework, this net effect β_1 will include both margins (i.e. (i) the average bribe amount changes as the set/type of people who collude changes and (ii) conditional on collusion, the bribe amount changes). To shed some light on these effects, in Section 6.2.1 we will also examine how the composition of those in the reassessed sample changes by estimating equation (5.5) on fixed characteristics of reassessed properties.

6 Results

In Section 6.1, we examine the impacts of the performance pay schemes on the key revenue and non-revenue outcomes of interest. Section 6.2 then probes the mechanisms through which changes in tax base occur in light of the model in Section 3. While we focus on the pay-for-performance aspect of the schemes, (i.e., price effects), Section 6.3 considers a variety of alternative explanations for the results, such as perceptions of additional monitoring, income effects, and interactions with supervisors. Section 6.4 concludes with a discussion of cost-effectiveness.

6.1 Main impacts

6.1.1 Impacts on Revenue Outcomes

Table 3 considers the impact of the performance pay schemes on (log) revenue at the end of each of the two years of the study. We first consider the impact on total revenue (columns 1 and 4). The remaining columns break this down into revenue derived from current year taxes and revenue from arrears. Current year revenue is about 5 times larger than arrears revenue. Arrears revenue is also substantially more variable over time, which is why the standard errors are larger when we examine arrears. Panel A reports the impact where we pool all three performance pay schemes and Panel B shows the impact for the schemes separately.

We find substantial impacts of performance pay on total revenue collected. Panel A, column 1 shows that compared to controls, revenue increased by 9.1 log points in treatment circles in the first year, and column 4 shows an increase of 9.4 log points in the second year. To interpret the magnitude of the effects, note that on average control circles experienced an increase in total revenue of about 25 log points between the baseline year and the end of the second year. Exponentiating, this implies that control circles grew by about 28% over the 2 years, and treatment circles grew by about 41%. Incentives thus led to a 13 percentage point increase in the growth rate, or a 46 percent higher rate of growth, over the 2 years of the experiment.

Examining the effects separately by current and arrears revenue, we find that the impact on current year revenue collection is 7.3 log points in Year 1 and 9.1 log points in Year 2. In contrast, there is a 15.2 log point increase in arrears revenue in Year 1, which falls to 11.3 log points (and is no longer statistically significant) in Year 2. Although these changes over the years are not statistically distinguishable, the point estimates suggest that inspectors, who exhausted much of the available pools of easily collectable arrears in the first year, switched their focus to increasing current year collection in the second year.

Separating the results by the three compensation schemes (Panel B), we see that, as one might expect, schemes that directly reward on revenue collection have a larger impact on revenue collected. Looking at current year revenue (where we have much more precise estimates for the aforementioned reasons), Column 5 shows that by the end of Year 2, Revenue circles collected 15.2 log points more revenue than control circles, compared to a 8.1 log point increase in Revenue Plus circles and a 3.5 log point increase in Flexible Bonus circles. When we test for equality of these coefficients we find that we can reject equality at the 10% level. Furthermore, when we test for equality between Revenue and an average of the multitasking schemes we are also able to reject equality (p-value 0.05). The magnitudes for the Revenue scheme are large: compared to the 39 percent average growth in current year revenue in control areas, revenue in Revenue circles grew by 62 percent. This implies that Revenue circles had a 58 percent (23 percentage point) higher growth rate in current revenue over 2 years than controls. The impact on total revenue collection–including arrears–was substantial as well: Revenue circles had 62 percent higher growth in than controls.

Our results show that performance pay schemes did lead to large increases in revenue, with schemes the rewarded explicitly/more on revenue collected seeing even larger increases. While our data verification checks gives us confidence that these schemes did in fact bring in real money, one potential concern is that these impacts might be due to temporary (and unreasonable) pressures put on taxpayers that could ultimately be undone through appeals (see, e.g. Das-Gupta and Mookherjee 1998). To investigate this we randomly sampled 22 circles, 2 (1 incentive, 1 control) in each of the 11 divisions, at the end of the second year of the experiment, and investigated all appeals that had been filed to date since the start of the experiment. We find that appeals are much too small (at most 1.5 percent of annual total revenues) to substantially change the results here, and find no economically meaningful or statistically significant differences in appeals rates or amounts between treatment and control areas.

6.1.2 Impacts on Non-Revenue Outcomes

To the extent that high powered incentives lead to excessive pressure to collect taxes and/or over-taxation/extortion, one may be concerned that the performance pay schemes – especially the Revenue only scheme – could adversely impact taxpayer satisfaction and assessment accuracy. Table 4 investigates these issues, and shows little evidence for such effects.

We examine the impact of the treatments on measures of taxpayer satisfaction and accuracy of tax assessment, using property-level survey data. Columns 1 and 2 in Table 4 examine the two measures of taxpayer satisfaction from the property survey in which we asked the respondent how they rated the "quality of service" of the tax department and how "satisfied" they were with their service (See footnote 9 for exact question wording). These are the exact measures which were incentivized in the Revenue Plus scheme, so it is instructive to examine not just whether they worsen in the incentive treatments in general, but whether the Revenue Plus scheme, and perhaps the Flexible Bonus scheme, mitigates this effect.

Panel A shows no statistically or economically meaningful treatment effect for either measure. In particular, on a 0-1 scale, the point estimates are -0.006 for quality of service and -0.011 for satisfaction, and we can reject a change in either measure of about 0.04 or larger.

Panel B examines the impacts separately for each scheme and finds the estimates for the Flexible Bonus are negative (-0.060 and -0.053 for quality and satisfaction, respectively), whereas the point estimates for Revenue Plus are positive (0.040 and 0.029, respectively). Although the results for each scheme are generally not statistically significant, it is worth noting that one can reject the null hypothesis of equality of the three schemes, or the null that that the Flexible Bonus scheme is equal to the other schemes. The estimates thus suggest that the Revenue Plus treatment, which explicitly incentivized quality and satisfaction, may have in fact led to higher levels of both compared to the Revenue and Flexible Bonus incentive schemes, though the magnitude of this impact is relatively small. The Flexible Bonus not only had the lowest performance in terms of revenue raised for the government, but also had worse outcomes on these other dimensions as well.

The zero average results on quality and satisfaction are quite robust. In particular, we also show in Appendix Table 4-G2 that the results are qualitatively unchanged if we use ordered probit models instead of the linearized variable with OLS or control for observable property characteristics (area, usage etc.).

In addition to these satisfaction measures, we examined other metrics that may reflect general attitudes towards the government, such as quality and satisfaction with other departments and stated preference for the incumbent party (based on self-reported voting behavior). These are shown in Appendix Table 13. In general, none of these metrics show meaningful differences between treatment and control. The only notable difference is that the pattern that Revenue Plus areas show higher satisfaction and quality of service appears generalized to other departments beyond just tax suggesting that there may be positive spillovers, which is consistent with citizens attributing a positive interaction in one government service to other related services.

Columns 3 and 4 in Table 4 examine the second main non-revenue dimension, the inaccuracy of tax assessment of the property. The results show no changes in inaccuracy or the tax gap overall (Panel A). When we explore the sub-treatments (Panel B), we do get some indication that Revenue Plus may have increased overall inaccuracy, although this does not seem to have an impact on the tax gap, which suggests that it may have raised both under and over-taxation for the full sample of properties. It is important to note, however, that this is the average effect for all properties. One potential reason we may not detect changes in this metric is that the number of properties affected may be small; we explore this in more detail when we focus on reassessed properties in Section 6.2.2 below.

On net, there are two key conclusions from the results thus far. First, compared to the control circles, we find that the incentives overall have a substantial, positive effect on revenue, with little detectable downside in terms of taxpayer satisfaction and the accuracy of tax perceptions for the typical property. Second, performance pay schemes with clearly defined objective criteria and with fixed proportional incentives tend to do better than more subjective, potentially uncertain, and multidimensional schemes. Comparing the Revenue and Revenue Plus scheme, we find that by year two the Revenue scheme had increased revenue by about 13 log points, whereas the Revenue Plus scheme increased current revenue by only about 9 log points; on the other hand, customer satisfaction appears slightly higher in the Revenue Plus scheme. The Flexible Bonus scheme did worse than either Revenue or Revenue Plus on all dimensions measured here. This provides suggestive evidence against subjective, potentially uncertain, and more multidimensional assessments and in favor of clearer, predictable, formula-based assessments that consider fewer metrics. This may be especially so in contexts where there may be concerns about credibility and how the more complex, subjective, and flexible assessments may be applied (see Baker et al. 1994, Prendergast and Topel 1996, Prendergast 1999, MacLeod 2003 for related theoretical work on subjective bonuses).

6.2 Changes in Tax Assessments and Rent Sharing

The model in Section 3 illustrates how taxpayers and tax collectors may collude to not pay taxes. The model shows how performance pay can make collusion harder and lead to higher tax collection and a switch from the collusive (high bribe, low tax) equilibrium to a non-collusive (low bribe, high tax) one, which could explain the increase in tax revenues. But the model also suggests that other taxpayers, who remain in the collusive equilibrium, would instead have to pay higher bribes to compensate tax inspectors for their for foregone incentive pay. This section explores these issues in more detail.

6.2.1 How many properties have valuation changed?

If bargaining breaks down, the theory suggests this should result in a change in the official tax valuation recorded by the government, τ. To explore changes in τ in the data, we examining impact the number and composition of properties whose official tax valuation was changed. We refer to these properties as "reassessed," and this includes both properties added to the official tax rolls for the first time as well as previously taxed properties whose tax valuation is updated.

Table 5 shows the total number of reassessed properties, broken down by properties reported as assessed for the first time and those who had previously been on the tax rolls but whose valuation was updated. For these data, we counted the number of properties added to the tax rolls or reassessed from the underlying tax registers, so this reflects the actual tax base as recorded by the government and is not subject to manipulation by tax staff in totaling. We control for the number of new and reassessed properties added in the baseline year (i.e. 2010-2011) to capture heterogeneity across circles in their underlying rate of change of properties.[27]

The results show a substantial increase in the number of properties whose valuations were changed in response to the treatment. On average (over the two year treatment period), there are 83 more properties per circle with new or updated valuations in treatment tax circles compared to controls, about an 86% increase over the control group. Most of this increase comes from properties that are newly *reported*. Column 2 shows that treatment circles add about 74 more newly valued properties to the tax rolls than controls (202% increase over the control group), while an additional 9 properties see their valuations updated. It is noteworthy however, that most of these properties are not actually new – 53 percent of these newly assessed properties were built before 2011 and a third constructed prior to 2006 (i.e. more than 5 years prior to the start of our experiment). In our field visits accompanying tax collectors, it was clear that they made visits to their tax circles frequently and were aware of where properties were located and their status (around two thirds of the supposedly new properties were within 500 meters of a property that reported having been visited by the tax collector). Therefore, as we discuss more later on, it seems more likely that the tax inspector was aware of these properties and they were strategically added to the rolls once performance pay incentives were introduced.

While these numbers document a substantial increase in activity in treatment circles compared to control circles in percentage terms, it is worth noting that the absolute numbers are still relatively small compared to the total number of properties in the circle: 74 new properties represents about 3 percent of the average number of taxable properties in the circle.[28] Nevertheless, as we explain below, these changing tax valuations are sufficient to explain essentially all of of the change in

[27]Note that since obtaining this data required a separate, detailed count of a different set of administrative records, we have this data only for a randomly-sampled set (approximately 50 percent) of circles.

[28]Note that these re-assessments are not necessarily exhausting a fixed supply of new or modified properties. Our property survey indicates that 1-2% of properties are built new each year, and an additional 2% have been either renovated or changed use in the past year, and since these properties tend to be larger and more valuable than the average property, there is substantial scope for additional ongoing increases in revenue collection on an ongoing basis.

revenue collected from the experiment. Increases in tax revenue can come through several margins: increasing the tax base by adding new properties to the tax rolls or updating their valuations, reducing exemptions granted to move tax bills closer to the gross tax base, or increasing the recovery rate of issued tax bills. Using the administrative data, we can decompose the increases in tax revenue into these three components. Doing so reveals that the vast majority of the increase in tax revenue is attributable to an increase in the tax base (see Appendix D), which are the types of reassessments documented in this section.

6.2.2 Changes in Collusion?

The model of collusion in Section 3 suggests that the treatment effects on taxes and bribes paid should have heterogeneous impacts among properties. For properties that switch from collusive to non-collusive equilibrium, we would expect to see an increase in taxes paid and a reduction in bribes. For properties that remain in the collusive equilibrium, we have more ambiguous predictions: the sum of bribes plus taxes paid should go up, but whether this comes from an increase in bribes, taxes, or some combination is less theoretically clear.[29] For properties that were in the non-collusive equilibrium before, and remain there, we would expect no changes.

To investigate these effects, in Table 6, Panel A we first estimate equation (5.4) in Section 5 on the general population of properties to capture how *typical properties* in treatment areas differ on these variables compared to equivalent properties in control areas. For the typical property we find that tax payments are essentially unchanged (Column 1) Note, though, that since the change in official revenues observed in the administrative data comes from a very small number of properties (as shown above), we would not necessarily detect it by looking across all properties, and indeed, we cannot reject the null of an average increase in taxes paid of the magnitude found in the administrative data.

In our model of collusive corruption, a low reassessment rate is consistent with many properties rebargaining bribes as a result of the incentive treatment. Indeed, columns 2 and 3 show that bribe rates - measured as the typical amount a property owner would pay in unofficial payments to the tax department over the course of the year for a similar property - increase substantially, by 594 Rupees (US $6, or about 32 percent higher compared to the average control area property).[30] The frequency of bribe payments also increases substantially. The one metric of corruption that does not change is the overall perception of corruption in the tax department.[31]

[29]Note that in the simple linear framework in Section 3 bribes unambiguously increase for properties that remain in the collusive equilibrium, but in the extension in Appendix A.1 with convex costs, the prediction on bribes becomes ambiguous.

[30]Note that the increase in average bribe payments comes entirely from the intensive margin, as we would expect from a shift in the collusive equilibrium. See Appendix Table 15 for more details.

[31]Note that we experimented in a pilot survey with asking directly whether the respondent had paid bribes. We experienced low response rates to this question, and found that respondents were much more forthcoming when we asked the question indirectly, i.e. what the going bribe rate was for a property that was "similar" to theirs. Note that this phrasing does not necessarily yield a precise average bribe paid, since respondents may answer the question either

Given limited statistical power in being able to detect the increase in overall taxes paid in treatment circles due to the lower frequency of reassessed properties in these circles, we now turn to specifically focusing our analysis on these reassessed properties. To examine this differential response for properties whose tax valuations were changed, we use data from the property survey and estimate equation (5.5). Recall that the property survey had both randomly sampled properties and those that were (over)sampled because they had been reassessed. In Panel B of Table 6, we estimate equation (5.5), which examines the differential impact between typical properties (i.e. those that stay in the same equilibrium they were in before), and reassessed properties (i.e. those whose tax bill changed, who may be disproportionately those that switch from one equilibrium to another). The coefficient β_2 from equation (5.5), i.e. the coefficient on $ReAssess$, captures how properties that are reassessed (or newly entered on the tax rolls) differ from the general population of properties in control circles, and β_1, the coefficient on $Reassess \times Treatment$, captures any additional difference in treatment circles (the treatment dummy is absorbed by the circle fixed effect).

There are several key results to note. First, compared with non-reassessed properties, properties in control circles whose valuations were changed pay substantially higher taxes – Rs. 2,763, or about 70 percent higher than the control group mean for random properties. This is even more true in treatment areas, where reassessed properties pay an additional Rs. 1,884 more than non-reassessed properties. The results here are consistent with the treatment effect on revenue we see in the administrative data. [32]On the other hand, the increase seen in bribes in treatment areas is not seen for reassessed properties, that is, the coefficient on $ReAssess \times Treatment$ is negative, and completely offsets the treatment effect for bribes on random properties shown in Panel A.[33]

---

conditional or unconditional on paying a bribe and the wording of the questions is not precise enough to reliably distinguish between the two. Since the frequency of bribes paid also goes up, however, this implies that even though we may not be able to estimate the precise magnitude, average bribe payments do in general increase.

[32]To see this, note that average tax in a circle is a weighted average of tax paid by reassessed and non-reassessed properties, i.e.

$$E[TaxPayment] = E[TaxPayment|Reassessed]P(Reassessed) + E[TaxPayment|NonReassessed]P(NonReassessed)$$

Based on our estimates here and data on reassessment rates (9% of taxable properties were reassessed in the cumulative two year treatment period in control circles and for simplicity we treat our general population sample as composed only of non-reassessed properties), this average in control areas is

$$(0.09)(3928 + 2763) + (0.91)(3928) = 4177$$

This gives an average tax per property of Rs. 4,177 in control areas. Using our treatment effect estimates (i.e. increases in the number of reassessed properties and the greater payments received from such properties and the effectively unchanged payments for non-reassessed properties), the analogous average tax in treatment circles is given by

$$(0.128)(3928 + 2763 + 1884) + (0.872)(3928) = 4523$$

An increase in the average tax per property from Rs. 4,177 to Rs. 4,523 represents a 8.3% increase in tax collection which is quite close to the observed effect from our admin data of over 9% (9.3 log points).

[33]Appendix Table 6-H repeats analysis of Table 6 broken down by the three subtreatments. The results do not show substantial differences in these dimensions among the three subtreatments.

Thus these results show, as suggested in the model, that performance pay for tax collectors leads to heterogeneous effects: increases in bribes for the majority of properties, but no increases in bribes with substantial increases in tax revenue for a small number of properties that switch from collusion to non-collusion. They also underscore that the increased revenue as a result of the performance pay schemes is on account of a small number of properties moving from a collusive to a non-collusive equilibrium and the corresponding substantial increase in taxes paid by such properties.[34]

Table 7 next examines whether there is an analogous differential response on non-revenue outcomes i.e. satisfaction, inaccuracy and the tax gap. The key results are for inaccuracy and the tax gap. Column 3 shows that reassessed properties are more accurate (i.e. less inaccurate) compared to non-reassessed properties. That is, there is a closer match between the tax liability computed by our independent surveyors and that computed by the tax department. Moreover, column 4 shows that while the typical (i.e. randomly-selected) property in the control group is under-taxed, this is eliminated in reassessed properties (i.e. adding the coefficient of 0.122 on re-assessment to the mean of -0.103 yields a net result of 0.019, which is not statistically significant from zero (p-value 0.191)); i.e. reassessed properties are on average taxed at the amount our independent survey team would predict. While these effects are similar in both treatment and control areas, they confirm the view of re-assessment as a bargaining breakdown: unlike typical randomly-selected properties, which in general are under-taxed, reassessed properties are assessed more accurately and are neither over- nor under-taxed on average.

It is also interesting to note that reassessed properties are not, broadly speaking, unsatisfied with the tax department. In fact, Table 7 shows that reassessed properties in general appear more satisfied with the tax department, and this is not different between treatment and control. One reason that there may be no change in satisfaction for these properties between treatment and control – even though they pay fewer bribes but much more taxes in treatment areas – is that the theory predicts that those who are reassessed and switch between the collusive and non-collusive equilibrium in response to the treatment are those who are closest to being indifferent between the two regimes. The switch from collusive to non-collusive equilibrium may therefore represent a second-order utility change for these property owners, even though it yields a first-order change in revenue for the government.

[34] While admittedly somewhat tentative, the results presented here help translate the changes in tax revenue we observed above into an elasticity of tax payments with respect to the bribe that one would need to pay to remain in the collusive equilibrium. In particular, note that, in the language of the model, the main tax estimates in Table 3 are estimating $\frac{d\log\tau}{dr}$, and the estimates in column 2 of Table 6 are estimating $\frac{db}{dr}$. We can thus compute the elasticity of tax payments with respect to the implied bribe price by calculating

$$\frac{d\log\tau}{dlogb} = \frac{\frac{d\log\tau}{dr}}{\frac{d\log b}{dr}} = \frac{\frac{d\log\tau}{dr}}{\frac{1}{b}\frac{db}{dr}}$$

Averaging across all schemes, $\frac{d\log\tau}{dr}$ is about 0.094, $\frac{db}{dr}$ is 594, and baseline b is on average 1875. This implies that $\frac{d\log\tau}{dlogb}$ is about 0.3.

6.2.3 Who gets reassessed?

If these re-assessments represent bargaining breakdowns, an interesting question is which property-tax collector pairs are affected. In the model, equation (3.5) shows that the increase in taxes comes from those properties on the margin of switching – i.e. those properties with taxpayer and tax collector values of the disutility of evasion parameters α and β such that they are close to indifferent being the high-bribe, low-tax equilibrium and the low-bribe, high-tax equilibrium.

To examine who these marginal properties are in the data, we consider how reassessed properties differ from typical properties, and how this differs in treatment versus control areas. The dependent variables are all captured from the property survey, and again we estimate equation (5.5) to capture the difference between reassessed properties and general properties, and how reassessed properties appear different in treatment areas relative to controls.

The results are presented in Panels A and B of Table 8, where Panel A examines characteristics of the property and Panel B examines characteristics of the owner. Panel A shows that reassessed properties are generally those (in both treatment and control areas) that are subject to higher tax rates than typical property. For example, reassessed properties have higher assessed values: according to the data we obtain from our independent survey, they have a Gross Annual Rental Value (i.e. tax base, before exemptions are applied) that is 67% higher than the mean property in control areas. They also have more floors, and are more likely to have been recently renovated, to belong to a more expensive tax bracket (tax category), to be commercial (which is taxed at a higher rate), and to be rented (which is again taxed at a higher rate).

Examining whether any of these margins change further in treatment circles, the point estimates suggest that, on net, reassessed properties in treatment areas have a GARV that is an additional *33*% larger (p-value of 0.21) than the average reassessed property in control areas. Therefore, reassessed properties in treatment areas have a 122% higher GARV than the typical property in control areas from the general population sample.[35] Incentivized staff also seem to focus more on commercial rented properties, which have the highest assessments per square foot of area. One interpretation is that commercial properties have a higher disutility from paying bribes (i.e. higher α) than residential properties and hence are more marginal.

Panel B considers differences in owner characteristics. One interesting finding is that those owners who report a close personal (family/friend) relationship with a politician are 1.3 percentage points (over a baseline value of 5.3% of properties so connected in control circles) less likely to be reassessed than typical properties. However, this effect is undone in treatment areas – so that while connected owners seem to enjoy an advantage in general, this no longer the case in treatment areas. We should caveat that this particular result be interpreted with caution, given that it is only one out of many coefficients examined. However it is interesting that a similar pattern holds for

[35] Another way to look at this is to plot, non-parametrically, the relationship between the probability of being reassessed and tax density, which is the tax valuation per unit of covered area. Appendix Figure F.1 shows that high tax density properties appear more likely to be re-assessed in treatment areas.

education: educated owners are in general more likely to be reassessed but this effect is undone in treatment areas. On net, the results in this section do paint a consistent picture: the performance incentives led inspectors to concentrate on a relatively small number of high-value properties.

6.3 Alternative Channels

6.3.1 Changes in collusion vs. greater inspector search effort?

We have interpreted our results so far in the context of changes in collusive behavior as a result of introducing the performance pay schemes. This is not meant to imply that changing effort on the part of tax staff might not play a role, and indeed reassessing taxpayers clearly requires some effort. Instead, we posit that we are unlikely to observe the above pattern of results without a change in collusion along the lines of our model.[36] That is, although a tax collector may respond to treatment by working harder at uncovering the true tax liability of a taxpayer (we assume this is known by the tax collector in our simple theoretical setup, but it may require effort to discover) or in getting recovery against that liability (as in standard moral hazard models such as Hölmstrom 1979), these forces alone are not consistent with the results we find on bribes.

Furthermore, evidence from self-reported behavior by inspectors does not seem to indicate such effort was important: In Appendix Table 16 we find little observable change in effort (total hours spent working per day etc.) reported by inspectors in treatment areas. The only result is that inspectors seem to be spending more time in the office and less in the field. While it is possible that time in the office is correlated with higher effort (e.g. filling out paperwork), it is not a priori what one would have expected in terms of effort, especially to the extent that the relevant margin was uncovering recent property changes. However, changes in collusion could quite plausibly imply more time in the office in order to change corresponding paperwork.

In addition, Appendix Table 23 presents further evidence to suggest that even for the properties that were newly added to the tax registers, the tax collector likely had prior knowledge about them. Specifically, we compare attributes of properties newly added to the register to attributes of properties that were verified as new based on the third party property survey. We find that the former are more likely to be rented and commercial, are larger in area, have a higher tax liability (though the results here are not significant), and have been occupied by the current tenant for longer. In fact, properties that are newly assessed are more likely to be rented (which is harder to observe without prior interactions), even conditional on other property characteristics. These results suggest properties newly added to the tax rolls are being systematically targeted, something that would not be feasible unless the tax collector had prior (private) information that is now being revealed.

All told, the results here paint a picture consistent with the theoretical framework: in pay for

[36]For example, to the extent that rebargaining also involves effort (e.g., due to increased paperwork), effort will increase mechanically under a collusive model as well.

performance regimes, most properties pay no more taxes but do pay somewhat higher bribes; but, some properties switch from the collusive to non-collusive equilibrium. Those properties that are reassessed do not experience the increase in bribes, but instead pay substantially higher taxes, are assessed more accurately, and are no longer under-assessed relative to what our independent survey reveals.

6.3.2 Mechanisms Beyond Price Effects

We have thus far interpreted our results as a result of changes in collusion due to the increased marginal incentives (i.e. price effects) provided to collect more taxes. However, the schemes we introduced also have other aspects that could also enhance performance. In this section we briefly discuss several of these alternate channels (additional details can be found in Appendix D). The objective of this section is not to definitely rule out these channels - since it is likely they do contribute to some extent - but rather to see how significant they may be. We conclude that while some of these channels may partly contribute, the price effects of the incentives still seem to be the primary way in which the incentives had an impact.

With any performance-based payment scheme that pays over and above baseline salary, two potential confounds are perceived monitoring effects and income effects. Perceived monitoring refers to the fact that the incentive scheme may affect agents' beliefs about how they will be monitored even outside of the explicit financial incentives. Income effects refer to the fact that a top-up incentive scheme increases the overall income of an agent, in addition to changing the price of the incentivized action. In this case, if honesty is a normal good (i.e. inspectors take bribes because they have a high marginal utility of income) or there are efficiency wage effects as in Becker and Stigler (1974), one could imagine that our effects are also due to income (and not just price) effects.

For both of these potential alternate channels, an important piece of evidence of the primary role of price effects is the difference between the three incentive schemes. In particular, all three schemes were equally salient to inspectors and all three performance-pay schemes were designed to generate approximately similar expected income (and indeed did so),[37] yet we saw above that they generate very different impacts on revenue: the increase in tax revenue in Revenue was almost double what was is in Revenue Plus, and the Flexible Bonus scheme produced no detectable tax impacts. These simple facts suggest prima facie that the different prices implicit in the different schemes are primarily what are driving the results, not the income transfer per se.

Several additional tests also confirm that price effects seem to be the primary explanation for our results. Starting in Year 2, we introduced an "information-only" scheme which provided identical information to the Revenue scheme, but with no financial payments. As discussed in Appendix E.1,

[37] Average payments to inspectors in Year 1 were: Rs. 255,608 in Revenue, Rs. 247,283 in Revenue Plus, and Rs. 297,370 in Flexible Bonus. Average payments in Year 2 were Rs. 255,773 in Revenue, Rs. 282,490 in Rev Plus, and Rs. 255,977 in Flexible Bonus.

this scheme produced no statistically significant effects on total or current revenue, and suggests that perceived monitoring explained no more than one-third the total impact. To investigate income effects, Appendix E.2 tests for any income effects using the fact that benchmarks in the Revenue and Revenue Plus schemes we determined based on the 2nd, 3rd, and 4th lags of revenue, but not the 1st lag, so the 1st lag of revenue generates a shock to income but not price effects. This approach also finds no strong evidence of income effects in this context.

Supervisory effort could also respond to the incentive scheme. While this is not a confound per se, it entails a slightly different interpretation of the results. However, we show in Appendix E.3 that a separate treatment scheme that explicitly rewarded supervisors (in an analogous manner to the Revenue scheme) had no effects on average, nor did it have substantial interactions with the direct incentives. We also show that inspectors do not report being more extensively pressured by their supervisors to work harder in areas with inspector incentives, so on net this does not appear to be an important part of the story.

Given that the randomization was conducted publicly (to ensure a perception of fairness), this meant that both control and treatment circles knew their respective identities. One potential concern is that control group inspectors may have become discouraged and performed worse, leading us to over-estimate treatment effects. While any such spillovers are less of a concern when comparing among the three treatment schemes, it is also worth noting that in general, the rate of growth of revenue in control circles during the experimental years was greater than during the previous years, suggesting that overall discouragement was not a first order concern. In addition, to test for spillovers more directly, Appendix Table 10 examines the impact of the treatment on nearby, neighboring control circles, where the treatment would be particularly salient, compared to control circles further away with whom inspectors interacted less often. If spillovers were to have occurred, they would likely have occurred locally, as inspectors in nearby circles share the same physical office space. We cannot reject the null of no spillovers.

Finally, one may also be concerned that the performance pay scheme could have increased inspectors' security of tenure within their circles. However, tax staff were told explicitly that normal transfer policy would be in effect during the study, and we show in Appendix Table 17 that there are no statistically significant differences in transfer rates among treatment and control staff. It is also worth noting that, again, this effect would be similar across schemes and not be responsible for the differences in performance we see in the three schemes.

6.4 Cost-effectiveness

From the government's and broader policy perspective, a natural question is whether these schemes were cost-effective, i.e. whether the additional revenue received in taxes exceeded the amount paid as incentives. As shown in Section 3, under the assumption that bribes represent a transfer rather than an efficiency loss, the change in net revenue for the government – revenue received in taxes less the amount paid in incentives – is a measure of the change in social welfare from the program.

For the Revenue and Revenue Plus scheme, which pay out to staff a percentage of revenue collected over a fixed benchmark, one would expect the net revenue to be positive so long as the benchmark was set sufficiently high that one is not paying out for infra-marginal collections. Of course, benchmarks cannot be set too high or else staff would not be in the money and would not be receiving incentives on the margin, so setting the benchmark is non-trivial. For the Flexible Bonus Scheme, the payments were fixed in advance, so it is less clear ex-ante whether net revenue for the government would be positive.

We focus on cost effectiveness in the second year of the program, when it was at scale. For each circle, we predict the revenue at the end of year 2 using our estimated treatment effects for each scheme.[38] We use the estimates to calculate the predicted additional revenue in treatment circles due to the treatment, and then sum this across treatment circles to obtain total additional revenue. The total costs are then simply the actual performance-based payments paid out under each of the schemes. Net revenue is the difference between predicted additional revenue and the incentive costs.

The results are shown in Table 9. Since the point estimates are slightly different depending on whether the information treatment is included as part of the control group (as in Table 3) or not (as in Appendix Table 3-I), we report the results both ways (Panel A and B respectively). Taken together the results show that net revenue is positive, so that the schemes are cost-effective. Dividing the net gain (revenue less costs) by costs to calculate a "return on investment" for the government shows a return of 15% (Panel A) to 30% (Panel B). For the Revenue scheme, which raised the most revenue, the return at the end of Year 2 ranges from 35% (Panel A) to 51% (Panel B). The Revenue Plus scheme earns 14% to 28% ROI, and the Flexible Bonus scheme loses money for the government.

Note that since a main channel seems to be an increase in net demand (i.e. new properties added to the tax rolls), to the extent these changes are permanent and last even after the treatments are discontinued, the long-run cost-effectiveness from a time-limited/temporary introduction of performance-based pay could be substantially higher than the numbers reported here.

7 Conclusion

Our paper examines the impact of introducing performance pay schemes in taxation. Taxation is interesting not only because it is quite feasible to design outcome-based pay mechanisms, but also because it presents interesting challenges in considering incentive pay mechanisms. In particular,

[38] The only change from our main specification is that we estimate reduced form treatment effects, where we weight each circle by the circle's revenue in the baseline year in order to account for any heterogeneity in treatment effects across circles of different sizes, which matters substantially for the impact on total revenue raised.

given the potential for collusion between the civil servant and the citizen, high-powered incentives are not simply about increasing worker effort to achieve the desired (i.e., incentivized) outcome. Instead, in such contexts incentives can increase the bargaining power of the civil servant with respect to the taxpayer, leading to potentially less desirable outcomes.

Our results suggest that, while these effects on bargaining are present, on net performance pay mechanisms can be quite effective in raising additional taxes, and that they can do so without generating too much animosity towards the tax department that was often associated with tax farming historically. While it is possible that such costs may show up over a longer than two years time-frame (though the concerns regularly expressed about raising tax rates suggest officials perceive these impacts to be fairly immediate), it is nevertheless instructive to examine why such costs might not be as high in our performance-pay schemes. In standard contract theory, a principal has to better incentivize an agent to the extent that the agent's objective function differs from the principal's. In taxation, to the extent that there is collusion - and our results suggest that this is an important margin - there is a clear wedge in such objectives in terms of raising taxes. Performance pay can therefore reduce this wedge by directly making the tax collector a (partly) residual claimant on taxes collected.

But what about divergences in political objectives between the politician/government and tax collectors? The historical tax farming literature suggests that collectors may have been less sensitive to political costs they imposed when raising taxes. However, tax collectors in our context may not be as free to raise taxes – they are not so locally powerful that they are unaffected by the displeasure of the population they tax. In fact, more often than not they may have weaker socioeconomic and political influence compared to those they are meant to tax, and so may also be quite concerned about the potential costs that raising excessive taxes may induce. Qualitatively, conversations with tax collectors suggested that this was a concern, i.e. the tax collector would justify lower collections by noting that the taxpayers could get them transferred or otherwise sanctioned both because the individual taxpayer may be quite influential and/or because they may collectively be powerful (e.g., shop-keepers' local associations). In fact, quite often (perhaps as a tacit means of justifying collusion) tax collectors would express sympathy to a taxpayers' unwillingness to pay taxes, particularly in poorer localities, given the general level of dissatisfaction taxpayers would have about how their taxes are utilized (locally) by the state.

So how might tax collectors balance their increased incentives to raise more taxes due to performance-pay schemes with a need to not increase taxpayer dissatisfaction? One could imagine two different types of potential responses. One response is to tax a large number of (poorer) property owners, who may have less influence or ability to push back, and to spare the more connected, wealthier owners of larger properties. Alternatively, inspectors could instead focus their efforts on a small number of high value owners. This would generate the largest return per property, and avoid alienating a large number of people, but could be risky if it alienates influential people. In a sense, this is a tradeoff between two types of influence: since each person gets one vote, smallholders have more votes per dollar, and hence more influence democratically, but largeholders may have more influence. The results here suggest that inspectors took the latter approach: focusing on a small number of high value property owners in a manner that didn't allow for "sub-coalitions" of these few to effectively form.

In terms of how collusion mediates the impact of performance pay, we find evidence that it indeed strengthens the bargaining power of the tax collector and that doing so has ambiguous effects on the incentivized outcome - the amount of tax collected. In fact, for the majority of taxpayers, their tax paid remains unaffected, although they end up paying higher rents to the tax collector as they re-bargain. While some taxpayers do end up paying more taxes and collusion breaks down, generating more revenue for the government, these results offer a word of caution that, unlike a world where the only margin is effort (and there is no collusion), simply introducing high powered incentives will not reduce the extent of inefficiencies. In fact, as our results show, for some subset of the population the amount of rents paid will increase. Thus if the goal is to both increase performance/collections and reduce rent-seeking, one may need to accompany a performance pay mechanism with stricter monitoring and direct penalties for rent-seeking.

Taken together, the results in our paper suggest that, notwithstanding historical concerns regarding tax farming and the relative absence of such high-powered incentives in developed economies, performance-pay schemes in taxation may be a promising avenue to explore for developing economies. The remaining question for governments is whether they can mitigate the potentially undesirable effects of the increased bargaining power tax staff have over taxpayers by more direct audit based processes that can effectively detect and penalize such collusion. The fact that our results show impacts on the tax base suggest that a promising direction may be to introduce high-powered incentives for short durations and at times when revealing information to the government is particularly important (such as when a major revaluation of properties or similar such reform is underway), and such schemes may need to be accompanied by complementary efforts at reducing corruption and better third party data verification processes. To the extent these concerns can be addressed, our results demonstrate that such schemes can be an important and financially and politically feasible way for emerging economies to undertake the essential and necessary task of raising tax revenue and enlarging their tax base.

References

Ashraf, N., Bandiera, O. and Jack, K.: 2013, No margin, no mission? a field experiment on incentives for public service delivery, *Unpublished Manuscript, London School of Economics* .

Bahl, R., Wallace, S. and Cyan, M.: 2008, Pakistan: Provincial government taxation, *Technical report*, International Center for Public Policy, Andrew Young School of Policy Studies, Georgia State University.

Baker, G., Gibbons, R. and Murphy, K. J.: 1994, Subjective performance measures in optimal incentive contracts, *The Quarterly Journal of Economics* **109**(4), 1125–1156.

Bartlett, B.: 1994, How excessive government killed ancient rome, *Cato J.* **14**, 287.

Becker, G. S. and Stigler, G. J.: 1974, Law enforcement, malfeasance, and compensation of enforcers, *The Journal of Legal Studies* pp. 1–18.

Besley, T. and McLaren, J.: 1993, Taxes and bribery: the role of wage incentives, *The Economic Journal* pp. 119–141.

Best, M. C., Brockmeyer, A., Kleven, H. J., Spinnewijn, J. and Waseem, M.: 2013, Production vs revenue efficiency with limited tax capacity: theory and evidence from pakistan.

Boycko, M., Shleifer, A. and Vishny, R. W.: 1996, A theory of privatisation, *The Economic Journal* pp. 309–319.

Carillo, P., Pomeranz, D. and Singhal, M.: 2014, Tax me if you can: Evidence on firm misreporting behavior and evasion substitution, *Technical report*, Harvard Kennedy School.

Chetty, R.: 2009, Is the taxable income elasticity sufficient to calculate deadweight loss? the implications of evasion and avoidance, *American Economic Journal: Economic Policy* **1**(2), 31–52.

City of Cambridge: 2014, *Annual Budget 2014-2015*.
URL: *http://goo.gl/rMTDjB*

City of Detroit, Michigan: 2013, Comprehensive annual financial report for the fiscal year ended june 30, 2013, *Technical report*.
URL: *http://goo.gl/5NjnIs*

Dal Bó, E., Finan, F. and Rossi, M. A.: 2013, Strengthening state capabilities: The role of financial incentives in the call to public service*, *The Quarterly Journal of Economics* **128**(3), 1169–1218.

Das-Gupta, A. and Mookherjee, D.: 1998, *Incentives and institutional reform in tax enforcement: an analysis of developing country experience*, Oxford University Press New York/Oxford.

Duflo, E., Greenstone, M., Pande, R. and Ryan, N.: 2013, Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from india, *The Quarterly Journal of Economics* **1499**(1545), 1499.

Feldstein, M.: 1999, Tax avoidance and the deadweight loss of the income tax, *Review of Economics and Statistics* **81**(4), 674–680.

Fisman, R. and Wei, S.-J.: 2004, Tax rates and tax evasion: Evidence from 'missing imports' in china, *Journal of Political Economy* **112**(2), 471–500.

Gertler, P. and Vermeersch, C.: 2013, Using performance incentives to improve medical care productivity and health outcomes, *Working Paper 19046*, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w19046*

Glewwe, P., Ilias, N. and Kremer, M.: 2010, Teacher incentives, *American Economic Journal: Applied Economics* **2**(3), pp. 205–227.
URL: *http://www.jstor.org/stable/25760225*

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

Hölmstrom, B.: 1979, Moral hazard and observability, *The Bell Journal of Economics* pp. 74–91.

Holmstrom, B. and Milgrom, P.: 1987, Aggregation and linearity in the provision of intertemporal incentives, *Econometrica: Journal of the Econometric Society* pp. 303–328.

Holmstrom, B. and Milgrom, P.: 1991, Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design, *JL Econ. & Org.* **7**, 24.

Kahn, C. M., Silva, E. C. and Ziliak, J. P.: 2001, Performance-based wages in tax collection: The brazilian tax collection reform and its effects, *The Economic Journal* **111**(468), 188–205.

Kleven, H. J., Kreiner, C. T. and Saez, E.: 2014, Why can modern governments tax so much? an agency model of firms as fiscal intermediaries, *Technical report*.

Kleven, H. J. and Waseem, M.: 2013, Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from pakistan*, *The Quarterly Journal of Economics* p. qjt004.

Kumler, T., Verhoogen, E. and Frï¿œas, J. A.: 2013, Enlisting employees in improving payroll-tax compliance: Evidence from mexico, *Working Paper 19385*, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w19385*

MacLeod, B. W.: 2003, Optimal contracting with subjective evaluation, *The American Economic Review* **93**(1), 216–240.

Mookherjee, D. and Png, I. P.-L.: 1995, Corruptible law enforcers: how should they be compensated?, *The Economic Journal* pp. 145–159.

Muralidharan, K. and Sundararaman, V.: 2011, Teacher performance pay: Experimental evidence from india, *Journal of Political Economy* **119**(1), 39–77.

Olken, B. A.: 2007, Monitoring corruption: Evidence from a field experiment in indonesia, *Journal of Political Economy* **115**(2), 200–249.

Olken, B. A. and Pande, R.: 2012, Corruption in developing countries, *Annu. Rev. Econ.* **4**(1), 479–509.

Parrillo, N. R.: 2013, *Against the Profit Motive: The Salary Revolution in American Government, 1780-1940*, Yale University Press.

Piracha, M. M. and Moore, M.: 2015, Understanding low-level state capacity: Property tax collection in pakistan, *Technical Report 33*, ICTD.

Pomeranz, D.: 2013, No taxation without information: Deterrence and self-enforcement in the value added tax, *Working Paper 19199*, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w19199*

Prendergast, C.: 1999, The provision of incentives in firms, *Journal of economic literature* pp. 7–63.

Prendergast, C. and Topel, R. H.: 1996, Favoritism in organizations, *The Journal of Political Economy* **104**(5), 958–978.

Rasul, I. and Rogger, D.: 2013, Management of bureaucrats and public service delivery: Evidence from the nigerian civil service, *Work. Pap., Univ. Coll. Lon* .

Shleifer, A. and Vishny, R. W.: 1994, Politicians and firms, *The Quarterly Journal of Economics* pp. 995–1025.

White, E. N.: 2004, From privatized to government-administered tax collection: tax farming in eighteenth-century france1, *The Economic History Review* **57**(4), 636–663.

World Bank: 2006, Proptery taxes in the punjab, pakistan, *Technical report*.
URL: *https://openknowledge.worldbank.org/handle/10986/8277*

World Bank: 2009, Government of the punjab property tax decentralisation program : Scope evaluation report, *Technical report*.
URL: *https://openknowledge.worldbank.org/handle/10986/12378*

Table 1: Experimental Design

	Randomization		Implementation	
	Year 1	Year 2	Year 1	Year 2
Revenue	53	72	47	68
Revenue Plus	54	74	48	68
Flexible Bonus	54	73	49	67
Information	0	70	0	66
Control	322	194	338	213

Notes: The first two columns (under Randomization) show the number of circles that were assigned to each of the three (or four) treatment types in each year. In cases where staff did not consent to treatment after the first ballot (in Year 1), circles were assigned treatment values of 1/3 for each main treatment type (i.e. Revenue, Revenue Plus, and Flexible Bonus). Values are rounded. The second two columns (under Implementation) show the number of circles that were actually implementing the treatment at the end of the fiscal year. Treatment wasn't implemented either because of lack of consent or because the initially selected circle staff were transferred to new posts. See text for more details.

Table 2: Summary Statistics

	Mean	SD	N
Panel A: Administrative Data			
Log Revenue (Total)	15.75	0.74	482
Log Revenue (Current)	15.52	0.73	482
Log Revenue (Arrears)	13.91	1.17	479
Log Tax Base (Total)	16.14	0.80	482
Log Tax Base (Current)	15.86	0.73	482
Log Tax Base (Arrears)	14.40	1.37	479
Log Non-Exemption Rate (Total)	-0.23	0.20	482
Log Non-Exemption Rate (Current)	-0.19	0.13	482
Log Non-Exemption Rate (Arrears)	-0.30	0.41	479
Log Recovery Rate (Total)	-0.16	0.18	482
Log Recovery Rate (Current)	-0.14	0.14	482
Log Recovery Rate (Arrears)	-0.19	0.29	479
Panel B: Survey Data			
Property successfully found in administrative records (dummy)	.84	.37	11,971
Quality of Tax Department [0-1]	.53	.22	6,050
Satisfaction with Tax Department [0-1]	.55	.23	6,050
Inaccuracy	.34	.27	9,870
Tax Gap	-.099	.42	9,870
GARV	32,302	252,426	10,787
Self-reported tax payment in FY 2013	3,562	18,604	12,000
Bribe Payment	2,073	3,932	5,993
Frequency of Bribe Payment	.76	.88	4,802

Notes: Panel A statistics from administrative data are shown at the end of Year 2 of the study (FY 2012-2013). Each observation is one of the 482 circles as defined at the time of randomization. Panel B statistics from the property survey are for properties from the random sample drawn from the field. The Inaccuracy and Tax Gap measures are available for only those properties that could be matched to the administrative records. Subjective variables - i.e., Quality, Satisfaction, Bribe Payment, and Frequency of Bribe Payment - are reported for circles from the first phase of the survey only (see Appendix B for more details).

Table 3: Impacts on Revenue Collected

	Year 1			Year 2		
	(1) Total	(2) Current	(3) Arrears	(4) Total	(5) Current	(6) Arrears
Panel A: Main Treatment						
Any treatment	0.091***	0.073***	0.152**	0.094***	0.091***	0.113
	(0.028)	(0.027)	(0.069)	(0.031)	(0.032)	(0.083)
Panel B: Subtreatments						
Revenue	0.118***	0.109***	0.134	0.129***	0.152***	0.005
	(0.035)	(0.034)	(0.099)	(0.043)	(0.044)	(0.133)
Revenue Plus	0.080	0.086*	0.072	0.093**	0.081*	0.175
	(0.053)	(0.052)	(0.110)	(0.045)	(0.049)	(0.114)
Flexible Bonus	0.071*	0.024	0.243**	0.056	0.035	0.148
	(0.038)	(0.035)	(0.098)	(0.041)	(0.042)	(0.108)
N	481	481	481	482	482	479
Mean of control group	15.671	15.379	14.030	15.745	15.518	13.915
Rev. vs. Multitasking p.	0.323	0.193	0.830	0.233	0.049	0.262
Objective vs. Subjective p.	0.530	0.090	0.212	0.220	0.084	0.634
Equality of Schemes	0.562	0.143	0.433	0.359	0.086	0.527
Joint significance	0.004	0.010	0.073	0.012	0.005	0.305

Notes: This table presents results on the impact of the performance pay schemes on revenue-based outcomes. We use instrumental variables regressions, where treatment status is instrumented with randomization results. The unit of observation is a circle, as defined at the time of randomization. Outcome variable is log revenue collection as of the end of the fiscal year, for total revenue (Columns 1 and 4), current year revenue (Columns 2 and 5), and collections against arrears (columns 3 and 6). Specification follows Equation 5.3 of the main text, and includes stratum fixed effects. 'Any treatment' in Panel A includes the 3 subtreatments in Panel B. The Information treatment is included in the control group. We report p-values from tests of equality of coefficients as follows: Rev. vs. Multitasking tests for equality between Revenue and the average of Revenue Plus and Flexible Bonus; Objective vs. Subjective tests for equality of the average of Revenue and Revenue Plus against Flexible Bonus; Equality of Schemes tests whether all coefficients are equal; and Joint significance tests joint null that all coefficients are equal to 0. Robust standard errors in parentheses. Standard errors are clustered by a robust partition of circles, i.e. the group of circles such that all circles that merged or split with each other are included within the same partition. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 4: Impacts on Non-Revenue Outcomes

	(1) Quality	(2) Satisfaction	(3) Inaccuracy	(4) Tax Gap
Panel A: Main Treatment				
Any treatment	-0.006	-0.011	0.004	0.007
	(0.022)	(0.022)	(0.012)	(0.022)
Panel B: Subtreatments				
Revenue	0.006	-0.006	0.002	-0.022
	(0.036)	(0.037)	(0.017)	(0.029)
Revenue Plus	0.040	0.029	0.028*	0.015
	(0.026)	(0.027)	(0.016)	(0.032)
Flexible Bonus	-0.060*	-0.053*	-0.016	0.029
	(0.031)	(0.032)	(0.018)	(0.031)
N	6050	6050	9870	9870
Sample	Phase 1	Phase 1	Full	Full
Mean of control group	0.538	0.555	0.339	-0.103
Rev. vs. Multitasking p.	0.683	0.876	0.813	0.159
Objective vs. Subjective p.	0.015	0.064	0.099	0.315
Equality of Schemes	0.014	0.059	0.090	0.344
Joint significance	0.035	0.129	0.160	0.533

Notes: This table presents results on the impact of the performance pay schemes on non-revenue outcomes. We use instrumental variables regressions, where treatment status is instrumented with randomization results. Unit of observation is a property. Specification follows Equation 5.5 of the main text, and includes stratum fixed effects. Quality and Satisfaction were measured on a 5 point Likert scale and re-scaled to a [0,1] interval. Tax Gap is the difference in the official gross annual rental value (GARV) minus our estimated GARV, divided by the sum of these. Tax Gap measures over/undertaxation, with positive coefficients indicating overtaxation. Inaccuracy is the absolute value of Tax Gap. Sample is restricted to Phase 1 of the survey for subjective outcomes (Quality and Satisfaction). The Information treatment is included in the control group. We report p-values from tests of equality of coefficients as follows: Rev. vs. Multitasking tests for equality between Revenue and the average of Revenue Plus and Flexible Bonus; Objective vs. Subjective tests for equality of the average of Revenue and Revenue Plus against Flexible Bonus; Equality of Schemes tests whether all coefficients are equal; and Joint significance tests joint null that all coefficients are equal to 0. Standard errors are clustered by robust partition of circles, i.e. the group of circles such that all circles that merged or split with each other are included within the same partition. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 5: Impacts on Number of Reassessed Properties

	(1) Total Number of Section 9 Properties Added to Tax Rolls in Treatment Period	(2) Number of New Properties Added to Tax Rolls in Treatment Period	(3) Number of Reassessed Properties Added to Tax Rolls in Treatment Period
Treatment	83.0*	74.0**	9.0
	(45.27)	(34.39)	(22.35)
N	234	234	234
Mean of control group	96.7	36.7	60.0

Notes: This table presents results on the impact of performance pay schemes on the number of properties that experience a change in tax status. Column 1 presents treatment effects on the total number of such properties added. The next two columns disaggregate this effect by whether the property is reported to have been previously registered on the tax rolls (Column 3) or not (Column 2). The sample consists of circles surveyed in Phase 2 (see text for details). Specification includes stratum fixed effects and controls for number of new and reassessed properties added in the pre-treatment (FY 2011) fiscal year. Standard errors are clustered by robust partition, the partition of circles such that all circles that merged or split with each other are included within the same partition. * p<0.10, ** p<0.05, *** p<0.01

Table 6: Impacts on Tax Payments and Corruption, by Reassessed Status

	(1) Self-reported Tax Payment	(2) Bribe Payment	(3) Frequency of Bribe Payment	(4) Perception of Corruption
Panel A: General Population Sample Only				
Treatment	-62.81	594.1*	.2021**	.0113
	(264.7)	(341.7)	(.0951)	(.0254)
N	11586	5993	4802	6050
Mean of control group	4069.425	1874.542	0.683	0.644
Panel B: Re-assessed and General Population Sample				
Re-assessed * Treatment	1884*	-557.4	-.1592*	-.0031
	(1083)	(380.1)	(.0942)	(.0221)
Re-assessed	2763***	-66.38	.0137	-.0191*
	(572.9)	(177.5)	(.0403)	(.0107)
N	16353	8207	6993	8268
Sample	Full	Phase 1	Phase 1	Phase 1
Mean of control group in gen. pop. sample	3928.252	1874.542	0.683	0.644

Notes: This table considers how the average property in treatment areas differs in terms of the tax payments and bribes it reports (Panel A) as well as asking whether these outcomes differ for reassessed properties (Panel B). In both cases we present instrumental variables regressions, where treatment status is instrumented with randomization results. Unit of observation is a property. Bribe Payment is the respondent's response to how much bribe they think others would pay for a similar property. Frequency of Bribe Payment and Perception of Corruption are graded on a 5 point rubric and scaled to the interval [0,1]. Panel A uses only properties from the random sample drawn from the field, while Panel B also includes properties that were selected from the official register of reassessments. The Re-assessed dummy in Panel B denotes such (reassessed) properties. The specifications in Panel A follow Equation 5.5 of the main text, with the exception of Column (1), which controls for self-reported baseline (FY 2011) tax payment. Specifications in Panel B follow Equation 5.6 of the main text. For Columns (2-4), sample is restricted to circles from the first phase of the survey (see text for details). In both Panels A and B, specifications include a control for whether the response came from the short version of the survey, and the phase of the survey (if applicable). The Information treatment is included in the control group. Robust standard errors in parentheses. Standard errors are clustered by robust partition of circles, i.e. the group of circles such that all circles that merged or split with each other are included within the same partition. * p<0.10, ** p<0.05, *** p<0.01

Table 7: Impacts on Satisfaction and Accuracy, by Reassessed Status

	(1) Quality	(2) Satisfaction	(3) Inaccuracy	(4) Tax Gap
Re-assessed * Treatment	0.009	0.005	0.001	-0.005
	(0.024)	(0.024)	(0.017)	(0.028)
Re-assessed	0.049***	0.044***	-0.061***	0.122***
	(0.013)	(0.013)	(0.009)	(0.015)
N	8268	8268	14173	14173
Sample	Phase 1	Phase 1	Full	Full
Mean of control group in gen. pop. sample	0.538	0.555	0.339	-0.103

Notes: This table examines whether non-revenue based outcomes differ for reassessed properties. The unit of observation is a property. Specification follows Equation 5.6 of the main text, and controls for whether the response came from the short version of the survey. Columns (1) and (2) restrict the sample circles from the first phase of the survey (see Appendix B for details). The Information treatment is included in the control group. Robust standard errors in parentheses. Standard errors are clustered by robust partition, i.e. the group of circles such that all circles that merged or split with each other are included within the same partition. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 8: Selection Effects on Reassessments

Panel A

				Components of GARV						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	GARV	Number of floors	Last renovation was ≤ 2 years ago	Land area (sq. feet)	Total covered area (sq. feet)	Main Road	Tax Category	Percent of property commercial	Percent of property commercial and rented	Tax Liability
Re-assess * Treatment	20137.796	0.002	-0.005	-32.599	852.092	-0.002	-0.226**	0.018	0.075**	3897.980
	(16187.550)	(0.050)	(0.020)	(82.473)	(771.516)	(0.048)	(0.088)	(0.037)	(0.029)	(3539.474)
Re-assess	24683.609***	0.078***	0.094***	37.396	-156.619	0.064***	0.212***	0.217***	0.176***	5503.481***
	(7944.915)	(0.026)	(0.011)	(57.199)	(379.299)	(0.024)	(0.044)	(0.019)	(0.015)	(1754.013)
N	15090	16352	16354	16352	16352	16352	15090	16226	16227	15090
Mean of control group in gen. pop. sample	36808.77	1.57	0.02	301.13	2779.82	0.46	3.78	0.35	0.17	6642.00

Panel B

	(1)	(2)	(3)	(4)	(5)	(6)
	Approximate age of owner	Owner's level of education	Per-capita wages	Predicted expenditure given assets	Connected to Politician	Connected to Politician/ Government/ Police
Re-assess * Treatment	-0.348	-0.523*	-821.749	110.798	0.021*	0.005
	(0.799)	(0.317)	(1078.191)	(213.234)	(0.012)	(0.027)
Re-assess	-0.656*	0.303*	13.126	-94.529	-0.013**	0.005
	(0.398)	(0.157)	(510.006)	(122.380)	(0.006)	(0.014)
N	13406	16254	13765	13954	16354	16354
Mean of control group in gen. pop. sample	50.70	9.19	16281.55	6292.58	0.05	0.36

Notes: Property level 2SLS regressions. Specifications follow Equation 5.6 of the main text, and includes a control for whether the response came from the short version of the questionnaire. Panel A looks at selection effects on property characteristics and Panel B looks at selection effects on owner/tenant characteristics. The characteristics in Panel A labeled Components of GARV are those that directly enter into the formula used to calculate GARV. Tax Category (Panel A, Column 7) is 7-tiered categorical variable with 7 being the most expensive tax bracket and 1 being the cheapest. Per-capita wages (Panel B, Column 3) is self-reported household expenditures divided by the total number of working household members. Predicted expenditure given assests (Panel B, Column 4) is the predicted value of a regression of household expenditure on series of dummy variables indicating various household assets. Standard errors in all panels are clustered by robust partition, the partition of circles such that all circles that merged or split with each other are included within the same partition. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 9: Cost-effectiveness of Incentives

	(1) Additional Revenue	(2) Cost of Incentives	(3) ROI
Panel A: Information in controls			
Any treatment	124,961,461	108,387,160	15.29
Revenue	50,578,024	37,349,784	35.42
Revenue Plus	40,671,290	35,549,342	14.41
Flexible Bonus	30,555,313	35,488,035	-13.90
Panel B: Information out of controls			
Any treatment	140,973,016	108,387,160	30.06
Revenue	56,269,064	37,349,784	50.65
Revenue Plus	45,539,845	35,549,342	28.10
Flexible Bonus	35,571,720	35,488,035	0.24

Notes: This table estimates the economic return generated by the performance pay schemes. Column 1 estimates the additional revenue due to treatment, calculated with a reduced form regression of log total revenue on log total baseline revenue, weighting observations by baseline revenue (in levels). For each treated observation, we generate a prediction of revenue collection under treatment and a prediction of revenue collection in absence of treatment and subtract to calculate the additional revenue due to treatment. The total additional revenue collection due to treatment is the sum of additional revenue collection across treated observations. Panels A and B show how the calculation changes depending on whether the Information treatment is included in the controls (Panel A) or dummied out (Panel B). Column 2 gives the actual costs of the incentive payments paid to circle staff under each scheme. Column 3 then presents Return on Investment (ROI), which is simply the percent increase in additional revenue above costs.