NBER WORKING PAPER SERIES

TAX FARMING REDUX:
EXPERIMENTAL EVIDENCE ON PERFORMANCE PAY FOR TAX COLLECTORS

Adnan Q. Khan
Asim I. Khwaja
Benjamin A. Olken

Working Paper 20627
http://www.nber.org/papers/w20627

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2014

This project is the results of collaboration among many people. We thank Jon Hill, Donghee Jo, 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. This RCT was registered in the American Economic Association Registry for randomized control trials under Trial number AEARCTR-0000252. The views expressed here are those of the authors and do not necessarily reflect those of the many individuals or organizations acknowledged here, nor those of the National Bureau of Economic Research.

Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors
Adnan Q. Khan, Asim I. Khwaja, and Benjamin A. Olken
NBER Working Paper No. 20627
October 2014
JEL No. D73,H26

ABSTRACT

Performance pay for tax collectors has the potential to raise revenues, but might come at a cost if taxpayers face undue pressure from collectors. 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 9.3 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 12.8 log points (62 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. Further analysis reveals that these 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 do report higher bribes. The results are consistent with a collusive setting in which performance pay increases collector's bargaining power over taxpayers, who either have to pay higher bribes to avoid being reassessed, or pay substantially higher taxes if collusion breaks down.

Adnan Q. Khan
London School of Economics
International Growth Centre
Houghton Street
London WC2A 2AE, UK
a.q.khan@lse.ac.uk

Asim I. Khwaja
Harvard Kennedy School
Harvard University
79 JFK Street
Cambridge, MA 02138
and NBER
asim_ijaz_khwaja@harvard.edu

Benjamin A. Olken
Department of Economics, E17-212
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
bolken@mit.edu

A randomized controlled trials registry entry is available at:
https://www.socialscienceregistry.org/trials/252
An Online appendix is available at:
http://goo.gl/1xifaf

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.

Historically, the state addressed this problem by rewarding tax collectors on the basis of tax collected. From the Roman Empire through the French Monarchy (Bartlett 1994, White 2004), regimes 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. But the empowered tax officials in these regimes were unpopular, and over the past several hundred years the world has moved increasingly to salaried tax officials (Parrillo 2013).

In developed countries with well functioning civil service systems, reliable financial and administrative data, and third party verification methods (Gordon and Li 2009, Kleven et al. 2011, Kleven forthcoming), there has been relatively little appetite for revisiting this question of compensation. But in developing countries, where tax inspectors are poorly compensated, corruption is considered rampant, and there is limited and only partial administrative and third party data to draw on (e.g., Carillo et al. 2014), the tradeoff between the expected benefits and costs of empowering tax officials is not so obvious. Therefore in such contexts with limited tax capacity mechanisms that may no longer be as desirable in developed countries may still offer a viable and effective solution. Not surprisingly, 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). Yet there is little rigorous evidence of the tradeoffs developing country governments face between increased revenue from incentives and the costs that may occur in the form of taxpayer dissatisfaction and over-taxation as a result of empowering tax collectors in this way.

In this paper, we provide what is to the best of our knowledge the first experimental evidence on this question. 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.[1] Many personnel in treated areas were able to

[1]The team of three tax personnel in each circle (an "inspector", "constable", and "clerk") together received either 20 percent, 30 percent, or 40 percent of all revenues collected above a benchmark. For equity reasons the percentage

double their baseline salaries or more through these incentives.

Given concerns about potential negative impacts of high-powered incentives on taxpayer satisfaction and assessment accuracy, the three schemes varied in both the extent to which they based performance pay explicitly on these non-revenue outcomes 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 (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.

It is important to note that the impact of even the revenue only scheme on tax revenues and corruption is not, ex-ante, obvious. 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 have to 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, high-tax equilibrium. As we subsequently show, performance pay could thus have heterogeneous effects on tax revenue and bribes among taxpayers.

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. Since tax assessment is determined formulaically from these property characteristics 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 9.3 log points based on the administrative data.

was pre-specified to be lower in larger circles, and the payments were divided among inspectors, constables, and clerks in a fixed proportion (40:30:30). Details are given in Section 4.1 below.

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 allows us to dig deeper by examining how being under a performance pay regime impacts taxpayers interactions with the tax department. We find heterogeneity in response among taxpayers that is consistent with there being collusion between tax inspectors and taxpayers. 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. 600 (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.

For the small number of properties whose tax valuation was formally changed (either newly assessed or re-assessed), 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 suggests 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 re-assessed 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) re-assessed 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 re-assessed properties have taxable value that is about 69 percent higher than the typical (non-reassessed) property. In treatment areas, the re-assessed properties are even more valuable than re-assessed properties elsewhere, by another 33 percent. Re-assessed 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 re-assessed 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 (34-50%), 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 (13-28%) 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.[2] 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

[2]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.

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). This paper shows 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 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 one might expect for the impact of 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). The tax is calculated by applying the appropriate multiplier to the size of the land and buildings on the property. 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 of this type of approach 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.[3] 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 2013).

[3] 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.

The lack of an explicit and transparent performance based reward system is 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 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 Conceptual Framework

Consider a simple setting where a taxpayer faces a true tax liability τ^*. The tax inspector knows τ^*, but can choose to report a lower tax liability to the government, τ, instead if he so chooses.[4]

[4]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. We give one example in Appendix

The tax inspector receives an incentive payment which is a constant fraction r of actual taxes paid, i.e. $r\tau$.

Both taxpayer and tax inspector face costs from colluding to report $\tau < \tau^*$. The taxpayer's cost of accepting a reduced tax liability is $\alpha(\tau^* - \tau)$ and the tax inspector's cost of giving a reduced tax liability is $\beta(\tau^* - \tau)$.

We assume that the taxpayer and the tax inspector engage in Nash bargaining, with the taxpayer potentially paying a bribe b as a transfer to tax inspector. If no agreement is reached, the taxpayer receives payoff $-\tau^*$ and the tax inspector receives payoff $r\tau^*$. If an agreement is reached, the taxpayer receives payoff $-\tau - \alpha(\tau^* - \tau) - b$ and the tax inspector receives payoff $r\tau - \beta(\tau^* - \tau) + b$.

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

$$\tau^* - \tau - \alpha(\tau^* - \tau) + r(\tau - \tau^*) - \beta(\tau^* - \tau) \tag{3.1}$$

which can be rewritten as

$$-\tau(1 - \alpha - \beta - r) + (1 - \alpha - \beta - r)\tau^* \tag{3.2}$$

This equation shows that if

$$r + \alpha + \beta < 1 \tag{3.3}$$

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

Suppose that γ is the bargaining weight of the taxpayer (and $1 - \gamma$ 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

$$b = [(\beta + r)(1 - \gamma) + \gamma(1 - \alpha)]\tau^*$$

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 $r + \alpha + \beta < 1$ and $\gamma < 1$, 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 $r + \alpha + \beta > 1$, then collusion disappears, bribes fall

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 (our property survey suggests the typical property is in fact under-taxed by around 30%) and we therefore do not incorporate it in the model. In addition, we simplify the model by treating the tax-collector as a single decision maker and not distinguishing between different members of the tax-collection team. While there could be interesting bargaining issues between these members, this is not something we could obtain any credible data on and also would unlikely generate qualitatively different insights.

to zero, and tax revenue increases from 0 to τ^*.

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 results here; 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 $\tau^* - \tau$, and are for at least one party strictly positive at $\tau^* = \tau$. [5]

It is also worth noting that while the costs are modeled in terms of deviations from the true tax liability, i.e. $\alpha(\tau^* - \tau)$, an alternative formulation would be to have costs in terms of bribes paid and received, i.e. to value bribes αb instead of b, 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 the in the model derived here if we specify the costs in terms of bribes rather than in terms of tax evasion.

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.

[5] In the linear model, conditional on colluding, one always sets $\tau = 0$, so the only reason taxes increase in this 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, and so τ could also change in response to r. In Appendix A.1 we allow for more general cost functions, imposing the restriction only that the marginal cost of reducing tax liability is always strictly positive for either the taxpayer or tax inspector (or both), and weakly monotonically increasing in the amount of tax evasion for both parties. 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. In the more general model, however, the results stem not just from individuals switching from full-collusion to non-collusion (which still occurs in the general model), but also because conditional on evasion continuing to take place, the level of tax evasion may decrease with r, with ambiguous implications for bribes.

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 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 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.

[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.

4.1.2 Revenue Plus

The Revenue Plus scheme was similar to the Revenue-based scheme, but included additional incentives to help address the multitasking problem inherent in the tax 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, 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, their 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.

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.[8] 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.[9]

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 talented employees based on all factors (including subjective ones) they observe.[10] In this

[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. These amounts are roughly three-fourths of typical inspector's and constable's/clerk's base pay.

[8] 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.

[9] 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.

[10] 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

treatment, staff were again divided into three groups (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 and pay was determined by group. Specifically, everyone in the treatment provisionally earned a base salary supplement roughly equal to their average salary.[11] 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.[12]

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 included (compared to Revenue Plus) 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 formula was 0.269 in Year 2.

subjective component of performance (Baker et al. 1994, Prendergast and Topel 1996, Prendergast 1999, MacLeod 2003).

[11] 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.

[12] 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.

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 of the Schemes

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.[13] 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.[14]

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.[15] 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 A.1 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), none is significant at the 10 percent level.[16]

[13] 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.

[14] 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.

[15] 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.

[16] 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 Online Appendix Tables we show that the main average effects of incentives do not seem to be driven by this one sub-treatment (Online Appendix E), and that controlling for the variables included in the balance table does not meaningfully change the results (Online Appendix F). The Online Appendices can be found at `http://goo.gl/1xifaf`.

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 Online 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),[17] it is still the case that the bulk of taxpayers do 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

[17] 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).

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 "re-assessed 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 re-assessed. 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.[18] 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.[19]

Corruption also appears to be prevalent. On average, respondents report that annual bribes paid for a property similar to theirs are around Rs. 2,000 (US $20) – about half of the amount they report

[18] 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.

[19] 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.

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.[20] 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.[21] 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.[22] 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 splits and merges since randomization.

[20] One might be concerned that the quality and satisfaction variables are simply picking up noise. However, Panel A of Appendix Table A.2 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 A.2 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.

[21] 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.

[22] Note 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. Online Appendix Tables 3-D through 8-D 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. The results are qualitatively similar.

To estimate the impact of the separate sub-treatments, we estimate the analogous regression separately by treatment:[23]

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

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 re-assessed, 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.[24] We include stratum fixed effects and cluster standard errors at the circle level. When available, we include controls for baseline level outcome variables.[25]

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

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

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 that unlike equation (5.4), we now include circle fixed effects (α_c) to capture fixed differences

[23] 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 $\beta_3 = \frac{\beta_1+\beta_2}{2}$).

[24] 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.

[25] 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 Online Appendix Tables 6-G and 7-G.

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 re-assessed 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_{ic} is the amount of bribes paid, the coefficient β_1 in equation (5.5) shows how the difference in bribes paid between re-assessed 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 6.3.1 we will also examine how the composition of those in the re-assessed sample changes by estimating equation (5.5) on fixed characteristics of re-assessed 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 decomposes the revenue impact into changes in the tax base, exemptions, and recovery rate. Section 6.3 then probes the mechanisms through which changes in tax base occur in light of the framework outlined in Section 3. While we focus on the pay-for-performance aspect of the schemes, (i.e., price effects), Section 6.4 considers a variety of alternative explanations for the results, such as perceptions of additional monitoring, income effects, and interactions with supervisors. Section 6.5 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 3.6 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 log points in treatment circles in the first year, and column 4 shows an increase of 9.3 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.

Thus 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 8 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 Online Appendix Table 4-B1 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 also 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 A.4. 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 re-assessed properties in Section 6.3.3 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, multidimensional incentives appear to matter. 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. This suggests that although the effects on taxpayer satisfaction in Revenue were small and not statistically distinguishable from controls, if one were worried about these issues, a multidimensional incentive scheme could potentially address them, albeit with substantial revenue costs. The Flexible Bonus scheme did worse than either Revenue or Revenue Plus on all dimensions measured here. This provides suggestive evidence against subjective and more multidimensional assessments and in favor of clearer, formulaic 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 Decomposing the Revenue Impact

To better understand the source of the revenue changes observed, we can decompose them further using the administrative data. There are three margins that could be affected: the officially assessed tax-base (before exemptions are granted), the amount of exemptions granted before issuing tax bills, and the amount of tax revenue collected conditional on the (post-exemption) tax base. These are related as follows:

$$Revenue = TaxBase * NonExemptionRate * RecoveryRate \tag{6.1}$$

where $NonExemptionRate = \frac{TaxBaseAfterExemptions}{TaxBaseBeforeExemptions}$ and $RecoveryRate = \frac{Revenue}{TaxBaseAfterExemptions}$. Taking logs of equation (6.1) we obtain an expression that additively decomposes the source of tax

revenue.

Table 5 reports the results of this decomposition exercise for both treatment years. Columns 1 and 5 begin by reproducing the same regressions of $\ln Revenue$ from Columns 1 and 4 of Table 3, with the addition of baseline $\ln NonExemptionRate$ and $\ln TaxBase$ as controls.[26] The results columns show that virtually all the impact is coming from changes in the tax base, particularly for current year revenue. This implies that re-assessments – which, recall, can be either finding new properties or changing the assessment on existing properties – are the main margin through which tax inspectors raise revenue.[27]

Online Appendix Table 5-C shows this decomposition separately for the three schemes. The only notable difference among the sub-treatments is that the Revenue treatment also shows a statistically significant impact on recovery rate in years 1 and 2, driven by the recovery rate in arrears. Thus, it appears that in the Revenue treatment, inspectors worked both on tax base and recovery rate (particularly for arrears and in the second year), whereas in the other treatments that had multidimensional incentives, inspectors focused more on the tax base, at least as we can measure it here.

6.3 Understanding Channels: Officially Assessed Tax Base Changes

The results above suggest that property tax revaluations, the mechanism by which the tax base is increased, are an important factor that contributes to the increased collections due to the performance pay schemes. These revaluations can take two forms: previously untaxed properties are added to the tax rolls for the first time, or previously taxed properties have their valuations are updated (and usually increased).

The conceptual framework in Section 3 illustrates how taxpayers and tax collectors may collude to not pay taxes, a likely scenario in the context we study. The framework 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. This section thus explores how the schemes shift this reassessment margin.

6.3.1 How many are re-assessed?

We begin by examining impact on the number and composition of properties that are re-assessed. Panel A of Table 6 shows the total number of re-assessed properties, broken down by properties reported as assessed for the first time and those who had previously been on the tax rolls but whose

[26] The addition of these controls both reduces the sample size slightly (due to incomplete baseline values) and also slightly changes the point estimates. The coefficients in Columns 2 to 4 (6 to 8) should add up to the coefficient in Column 1 (5)

[27] It is also interesting to note that the tax base for arrears increases in year 1 but not in year 2. The arrears tax base can only increase if there are collections not made the year before, or if past uncollected amounts are now added in due to a valuation adjustment that is retroactively applied. Given the performance pay incentives in year 1, therefore, by year 2 there is likely not much room left to improve the arrears tax base.

valuation was updated. For these data, we counted the number of properties added to the tax rolls or re-assessed 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 re-assessed properties added in the baseline year (i.e. 2010-2011) to capture heterogeneity across circles in their underlying rate of change of properties.[28]

The results show a substantial increase in the number of re-assessments. 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. Note that most of these properties are not in fact new – 53 percent of these newly assessed properties were in fact built before 2011, and the average year of construction was 2006. 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. 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.

6.3.2 Who gets re-assessed?

To examine how those properties that are reassessed differ from typical properties, 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 sampled because they had been re-assessed. The coefficient on *Reassess* shows how re-assessed properties differ from the average property in control circles; the interaction term $Reassess \times Treatment$ captures how this difference between re-assessed and non-reassessed properties differs in treatment areas relative to controls. Note that the *Reassess* dummy refers to all properties whose valuation was changed, including both properties reported on the tax rolls for the first time and those whose valuations were updated. The dependent variables are all captured from the property survey.

The results are presented in Panels B and C of Table 6, where Panel B examines characteristics of the property and Panel C examines characteristics of the owner. Panel B shows that re-assessed properties are generally those (in both treatment and control areas) that are subject to higher tax rates than typical property. For example, re-assessed 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 69% higher than the mean property in control areas. They also have more floors, and are more likely to have been recently renovated,

[28] 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.

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, re-assessed properties in treatment areas have a GARV that is an additional *33*% larger (p-value of 0.21) than the average re-assessed property in control areas.Therefore, reassessed properties in treatment areas have a 127% higher GARV than the typical property in control areas from the general population sample. Interestingly, the focus on properties in higher tax categories (i.e. that are taxed at a higher rate, conditional on property type) is undone in treatment areas, suggesting that tax staff generally give lower category properties a pass until incentives make it worth their while to do otherwise. Incentivized staff also seem to focus more on commercial rented properties, which have the highest assessments per square foot of area.

Panel C 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 re-assessed than typical properties. However, this effect is undone in treatment areas – so that while connected owners seem to enjoy an advantage in general, this is undone 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. A similar pattern holds for 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 paint a consistent picture: the performance incentives led inspectors to concentrate on a relatively small number of high-value properties, likely those that were the most lucrative or represented the next feasible ones to focus on.

6.3.3 Do the Reassessed Pay Differently?

The theoretical framework 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 heterogeneous effects, in Table 7, 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 average property we find that tax payments are essentially unchanged (Column 1). This is nevertheless

[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.

consistent with the revenue impact observed in the administrative data: Since only about 9 percent of randomly selected properties have been re-assessed over the treatment period, if the tax increase in treatment areas stems from re-assessed properties, then we do not have sufficient sample size in the survey data to detect the increase we observed in the administrative data. Bribe payments and frequency, on the other hand, 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 584 Rupees (US $6, or about 30 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.

In Panel B of Table 7, we estimate equation (5.5), which examines the differential impact between the typical properties (i.e. those that stay in the same equilibrium they were in before), and re-assessed 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 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, re-assessed properties in control circles pay substantially higher taxes – Rs. 3,430, or about 73 percent higher than the control group mean for random properties. This is even more true in treatment areas, where re-assessed properties pay an additional Rs. 2,248 more than non-reassessed properties.[31] On the other hand, the increase seen in bribes treatment areas is not seen for re-

[30]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 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.

[31]The results here are consistent with the treatment effect on revenue we see in the administrative data. 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

$$(0.09)(4713 + 3430) + (0.91)(4713) = 5022$$

This gives an average tax per property of Rs. 5,022 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

$$(0.128)(4713 + 3430 + 2248) + (0.872)(4713) = 5440$$

An increase in the average tax per property from Rs. 5,022 to Rs. 5,440 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).

assessed 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.[32] Thus these results show, as suggested in the conceptual framework, that a key margin is whether incentives lead to changes in official assessments, with very different outcomes for those that stay in the collusive equilibrium and those that do not. 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.

Table 8 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 re-assessed 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. re-assessed 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, re-assessed properties are assessed more accurately and are neither over- nor under-taxed on average.

It is also interesting to note that re-assessed properties are not, broadly speaking, unsatisfied with the tax department. In fact, Table 8 shows that re-assessed 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 re-assessed 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.

6.3.4 Changes in collusion vs. greater inspector 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. However, even in the absence of any collusion, one could

[32] Online Appendix Table 7-C repeats analysis of Table 7 broken down by the three subtreatments. The results do now show substantial differences in these dimensions among the three subtreatments.

find results if inspectors simply worked harder in 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). While the results we find on bribes would not be consistent with a model where there is no collusion, evidence based on self-reported behavior by inspectors does not seem to indicate such effort was important: In Appendix Table A.5 we find that 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.

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

6.4 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. First, inspectors in these schemes receive more salient information and may be perceived as facing greater monitoring. Second, inspectors receive more income in the treatments than in the control, so there are potentially income effects as well as price effects. Third, supervisors might change their behavior in response to the incentives. While these effects do not raise concerns in terms of identifying the causal effect of the pay for performance scheme per se, unpacking them can help offer a more nuanced understanding of the various mechanisms and channels that may contribute to the results we observe. 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.

6.4.1 Information and monitoring effects

As part of the schemes, each quarter all tax staff received a form explaining how their payments were calculated, which showed their revenue collected and the corresponding benchmarks. While

this is not new information for them (it is calculated based on historical information they themselves reported), it was presented in a slightly different and potentially more salient way, comparing performance explicitly against historically derived benchmarks. Moreover, the fact that inspectors were called in for a group meeting each quarter and received printouts of their performance could have also created a sense of salience and/or being more intensively monitored.

There are several reasons to suspect that these type of information and monitoring effects do not primarily drive the results. First, they were identical in the three schemes – each quarter inspectors in each scheme received the same information sheets at quarterly meetings, yet the three schemes had substantially different results, to the point where the flexible bonus scheme had no detectable impact on revenue. This suggests that if information and monitoring effects were present, they were not accounting for the bulk of the impacts we find.

Second, to examine these effects more directly, starting inYear 2 we introduced an "information-only" scheme. In this scheme inspectors received the same type of training and quarterly information sheets (with benchmarks) as inspectors in the Revenue schemes, but without any additional financial compensation. This scheme thus nets out the effects from the actual incentive payments from any other effect that may be in the other schemes.[33]

While in our baseline specifications we included these circles as part of the control group, in Panel A of Table 9 we separate out the information scheme and compare it to the control group. The results in Table 9 show generally positive point estimates associated with the information scheme (particularly in the arrears treatment), but for total and current year revenue they are not distinguishable from zero. For current-year revenue, the point estimate is that the information scheme is associated with 7.1 log points higher revenue, compared with 16.8 log points for the Revenue scheme (a test for equality has a p-value of 0.093). So while it does seem like there may have been some effect, likely coming from the sense of being monitored, the large (and likely more sustained) fraction of the effect of the Revenue scheme is attributable to the financial aspect of the performance incentives.

6.4.2 Income effects

To the extent that 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),

[33]It is possible that this scheme created an additional "anticipation" effect that was not present in the other schemes. While we had made clear that this scheme did not provide any payments (and as we saw earlier in Table A.3 inspectors generally correctly identified which scheme they were in), there is some evidence to suggest that information only selected inspectors may have thought that they were more likely to be selected for a future performance pay scheme. In Online Appendix Table 13 we first confirm (column 1) that inspectors in the information scheme were not more likely than inspectors in control circles to believe they would receive any monetary or non-monetary rewards from the scheme. However, while they did not give a systematically higher chance of being selected in a subsequent years ballot for a performance pay scheme (column 2), when explicitly asked to compare their chances of being selected compared to control circles or current treatment circles (columns 3 and 4), they gave higher odds suggesting they (incorrectly) felt they had a higher chance of future selection.

one could imagine that our effects are also due to income (and not just price) effects.

There are several pieces of evidence that suggest that the impacts observed are not being driven by income effects. First, all three of the performance-pay schemes were designed to generate approximately similar expected income (and indeed did so),[34] 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.

In addition we can directly test for any income effects by taking advantage of 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.[35] Since revenue can be closely approximated by an AR(1) process, this suggests a way of identifying income effects.[36] Specifically, we regress

$$LnRevenue_t = \gamma_1 LnRevenue_{t-1} + \gamma_2 LnBenchmark_{t-2} + \epsilon \tag{6.2}$$

to form the prediction $\hat{LnRevenue_t}$, and then exponentiate to get $\hat{Revenue_t}$ and $\hat{Benchmark_{t-2}}$. This is the amount of revenue that would be collected under business-as-usual. An inspector in the Revenue or Revenue Plus group would therefore expect to earn

$$IncomeShock_t = \alpha \left(\hat{Revenue_t} - Benchmark_{t-2} \right) \tag{6.3}$$

simply from business-as-usual. Since there is heterogeneity across inspectors in $IncomeShock_t$ (due to idiosyncratic variation in $Revenue_{t-1}$ conditional on $Benchmark_{t-2}$), this identifies the pure income effect that an inspector randomized into Revenue or Revenue Plus would receive compared to an inspector in the control group.[37] Since $IncomeShock_t$ is defined in both treatment

[34] 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.

[35] The reason for this was both due to design and logistical considerations. In terms of design, not having the previous year's performance be part of the benchmark helped lessen ratchet effects in subsequent years (i.e. doing well in year 1 did not mean benchmarks for year 2 were higher). Logistically, benchmarks for 2012-2013 needed to be announced by the second week of July 2012, but the 2011-2012 revenue collection data would not be fully compiled, data-entered, and cleaned until August 2012. This meant that 2011-2012 revenue collection data could not be used in the computation of benchmarks for the 2012-2013 fiscal year.

[36] Specifically, if one regresses log revenue on its first 4 lags, the first lag has a coefficient close to 1 with an F-statistic of 314.4; the remaining 3 lags together have a joint F-statistic of only 5.2. We should note that this does not mean of course that benchmarks based on the 2nd through 4th lags are meaningless, just that the first lag is close to a sufficient statistic. For example, if one regresses current revenue on the 2nd, 3rd, and 4th lags only (omitting the first lag), one obtains an F-statistic of over 1,000; it is only once one also includes the 1st lag that the remaining lags have little explanatory power.

[37] In practice, this calculation is slightly more complicated, since in the first year of the schemes there were separate benchmarks for current-year revenue and arrears revenue (in year 2 they were combined). In year 1 we therefore estimate separate income shocks using equations (6.2) and (6.3) for current and arrears and then add to get the total income shock. Note also that this only works for the first year of the program (2012); in the second year, the first

and control areas, we can interact treatment status with the income shock that they would receive if they received a treatment to identify the treatment effect. Specifically, we estimate:

$$\begin{aligned} LnRevenue_{ct} &= \beta_1 Treatment_c + \beta_2 LnIncomeShock_{ct} \times Treatment_c + \beta_3 LnIncomeShock_{ct} \\ &\quad + \beta_4 LnRevenue_{ct-1} + \beta_5 LnBenchmark_{ct-2} + \epsilon \end{aligned} \tag{6.4}$$

The key coefficient of interest is β_2.[38]

Panel B of Table 9 presents the results. There is no difference in performance based on the infra-marginal component of the revenue treatments, i.e. there is no evidence of any income effect. Again, this suggests that the key component is the price effect, not the income effect.[39]

6.4.3 Supervisory incentives

To the extent that circle staff were also aided by their supervisors, one could examine this effect by directly examining the impact of performance-pay for supervisory tiers. Starting in Year 2 of the experiment, such supervisory performance-pay was also introduced. These rewards were very similar to the Revenue incentive treatment, but they were paid based on the average performance above benchmarks for all of the taxable units (and hence staff) under their supervision.

There are two levels of supervisors – Excise and Taxation Officers (ETOs) and Assistant ETOs. We randomized at the level of ETO and treated all AETOs who worked underneath them. Supervisors are not only responsible for monitoring the performance of the field staff and ensuring that collection targets are being met, but they can also directly aid in the collection process, especially in terms of supporting and imposing stronger sanctions on non-taxpayers and in handling appeals. All ETOs and AETOs had a mix of treatment and control circles working beneath them. Note that since we randomized at the level of 51 ETOs, we report randomized-inference based p-values, which are accurate in small samples and accounts for the clustering of the randomization at the ETO level.[40] With only 51 ETOs randomized in this treatment, compared to almost 500 circles in the main experiment, the level of statistical power is much lower here, but the Tax department wanted to include this scheme nonetheless, especially given the success of the circle staff schemes

lag of revenue is the revenue realized in the first treatment year, which is endogenous.

[38] Note that $IncomeShock_{ct}$ is not quite a linear combination of $Revenue_{ct-1}$ and $Benchmark_{ct-2}$, given the exponentiation and subtraction, so we include the main effect of this as well.

[39] An alternate approach would be to directly examine price effects using the fact that different circles received different incentive rates α. In principle, since the reward rate changes discontinuously at the 50th and 75th percentile of baseline circle size, one can apply RD techniques to estimate the impact of a higher reward rate. The challenge is power, as the number of circles close to the discontinuity is very small. When we apply this approach, we find positive but noisy estimates of being in the 30% or 40% reward rate compared to the 20% reward rate in both years, though the results in 2013 are somewhat sensitive to the functional form used for the running variable (results available on request). However, the standard errors on these estimates are quite large (around 0.12 log-points for current and 0.35 log-points for arrears).

[40] We found through Monte-Carlo simulations that conventional cluster-robust standard errors appear too small in this context and over-reject the null.

in the first year.

Panel C of Table 9 reports the results of the supervisory scheme. The unit of observation remains a circle. We find no effect – the point estimates are in fact negative for total, current, and arrears revenue, though they are never statistically distinguishable from zero. We have further investigated whether there are interactions between supervisors and the staff under them – i.e. is there a particular synergy from having both supervisor and staff incentivized, or are the effects orthogonal to one other. The results suggest that, if anything, paying supervisors only may in fact be detrimental to overall collections, though this is only marginally statistically significant and only for current-year revenue (See Online Appendix Table 11). We should caution that, given the imprecise estimates on the supervisory treatments due to lower sample size, it may be prudent not to make too much of this effect.

We also examine whether inspectors in treatment schemes believed they were being pressured more extensively by their supervisors to work harder. The results are shown in columns 1 and 2 of Online Appendix Table 12, and show that that there is, on average, no difference in perceived pressure from supervisors between treatment and control areas. On net, the results presented here suggest that increased pressure from supervisors does not appear to be an important part of the channel driving the effects, both because rewarding supervisors directly has little effect and because we do not find any reported increase in supervisory pressure/support when tax circle staff are incentivized.

6.4.4 Treatment Spillovers

While it was essential for fairness to conduct the randomization publicly, 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. When comparing among the three treatment schemes, all of whom were treated and which showed very different effects, such spillovers are less of a concern. For control circles, Online Appendix Table 14 tests for the presence of spillover effects by examining 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. We cannot reject the null of no spillovers. In fact, the small, positive point estimates suggest that, if anything, control circles tend to perform weakly better when a greater fraction of their neighbors are treatment circles, suggesting that if anything our estimates are slight under-estimates, rather than over-estimates, of the true treatment effect.[41]

[41] The weak positive effect holds regardless of the radius used to identify neighbors, or the particular cutoff point chosen for defining a "spillover control" (i.e. a control that has a relatively large fraction of its neighbors being treatment circles).

6.5 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.[42] For the Revenue and Revenue Plus scheme, which pay out to staff a percentage of revenue collected over a fixed benchmark, one would expect them to be cost-effective 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 low 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 it is cost effective or not.

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.[43] 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.

The results are shown in Table 10. 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 Online Appendix Table 3-D), we report the results both ways (Panel A and B respectively). Taken together the results show that the schemes are cost-effective, with additional returns exceeding costs. Dividing the net gain (revenue less costs) by costs to calculate a "return on investment" for the government shows a return of 14% (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 34% (Panel A) to 50% (Panel B). The Revenue Plus scheme earns 13% 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

[42] Under the assumptions that there is no dead-weight loss from additional payments (i.e. additional property tax payments are non-distortionary, conditional on statutory tax liability), and that the social cost of any additional effort exerted by tax inspectors is negligible, the program being cost-effective in a revenue sense implies that it increases social welfare. To see this, assume the government has fixed expenditure needs and assume that any additional revenue raised here reduces the amount that needs to be raised by some distortionary tax; or, alternatively that the government spends each additional dollar it has with some social benefit greater than 1 (if social benefits were less than 1 it would not be raising taxes).The social welfare gain is thus the net change in revenue to the government (i.e. additional revenue less expenditures on incentives) adjusted by the social loss from the avoided distortionary tax or social gain from induced additional spending. If collecting more property taxes holding the official rates fixed is distortionary, or if the social losses from higher induced efforts from tax collectors are non-trivial, then social welfare gain will be lower than cost effectiveness.

[43] 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.

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 two interesting challenges in considering incentive pay mechanisms: First, there is a danger of "over-incentivizing" the civil servant. In the context of taxation, as with historical tax farming, high-powered performance pay can potentially lead to extortion and excessive pressure on the taxpayer and ultimately lead to citizen discontent, an undesirable outcome for the state. Second, taxation naturally brings with it the potential for collusion between the civil servant and the citizen. Thus high-powered incentives are not simply about increasing worker effort to achieve the desired (i.e., incentivized) outcome, but incentives in such contexts can empower the civil servant in a way that impacts the bargaining between the civil servant and citizen, and can increase the bargaining power of the civil servant with respect to the taxpayer leading to potentially less desirable outcomes. Our results shed novel light on both these aspects.

In terms of the trade-off between revenue and non-revenue costs of high powered incentives, we find that 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 the tax collector? The historical tax farming literature suggests that tax collectors may have been less sensitive to the 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. To the extent that this is the case, they 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 the presence of collusion mediates the impact of performance pay, we find compelling 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, as we note above, 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 amounts of rents paid will increase. Thus if the goal is to both increase performance/collections and reduce rent-seeking, one may need to accompany 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.

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.

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.

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.: forthcoming, How can scandinavians tax so much?, *Journal of Economic Perspectives* .

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S. and Saez, E.: 2011, Unwilling or unable to cheat? evidence from a tax audit experiment in denmark, *Econometrica* **79**(3), 651–692.

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. and Moore, M.: 2013, The informality of formal tax collection in pakistan.

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* .

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	0.81	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)	0.84	0.37	11,971
Quality of Tax Department [0-1]	0.53	0.22	6,050
Satisfaction with Tax Department [0-1]	0.55	0.23	6,050
Inaccuracy	0.34	0.27	9,879
Tax Gap	-0.099	0.42	9,879
GARV	31,915	248,026	11,186
Self-reported tax payment in FY 2013	4,246	20,255	10,047
Bribe Payment	2,073	3,932	5,993
Frequency of Bribe Payment	0.76	0.88	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	0.090***	0.073***	0.152**	0.093***	0.091***	0.113
	(0.028)	(0.027)	(0.069)	(0.031)	(0.032)	(0.083)
Panel B: Subtreatments						
Revenue	0.117***	0.109***	0.134	0.128***	0.152***	0.005
	(0.035)	(0.034)	(0.099)	(0.044)	(0.044)	(0.133)
Revenue Plus	0.080	0.086*	0.072	0.092**	0.081*	0.175
	(0.053)	(0.052)	(0.110)	(0.045)	(0.049)	(0.114)
Flexible Bonus	0.070*	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	15.672	15.379	14.030	15.745	15.518	13.915
Rev. vs. Multitasking p.	0.322	0.193	0.830	0.237	0.049	0.262
Objective vs. Subjective p.	0.530	0.090	0.212	0.222	0.084	0.634
Equality of Schemes	0.561	0.143	0.433	0.363	0.086	0.527
Joint significance	0.004	0.010	0.073	0.014	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. 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	0.006	0.006
	(0.022)	(0.023)	(0.012)	(0.022)
Panel B: Subtreatments				
Revenue	0.006	-0.006	0.005	-0.027
	(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	9879	9879
Sample	Phase 1	Phase 1	Full	Full
Mean of control group	0.538	0.555	0.339	-0.103
Rev. vs. Multitasking p.	0.681	0.875	0.973	0.120
Objective vs. Subjective p.	0.014	0.061	0.081	0.280
Equality of Schemes	0.014	0.060	0.097	0.276
Joint significance	0.035	0.130	0.162	0.457

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. 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 Tax Base and Recovery Rates, All Treatments

	Year 1				Year 2			
	(1) Revenue	(2) Tax Base	(3) Non-Exemption Rate	(4) Recovery Rate	(5) Revenue	(6) Tax Base	(7) Non-Exemption Rate	(8) Recovery Rate
Total								
Any Treatment	0.075***	0.089***	-0.025	0.011	0.089***	0.053	0.006	0.029
	(0.027)	(0.029)	(0.018)	(0.023)	(0.030)	(0.033)	(0.020)	(0.019)
Curent								
Any Treatment	0.073***	0.084***	0.000	-0.012	0.096***	0.067**	0.018	0.011
	(0.028)	(0.028)	(0.014)	(0.022)	(0.032)	(0.030)	(0.016)	(0.016)
Arrears								
Any Treatment	0.111*	0.133*	-0.053	0.032	0.075	-0.006	0.053	0.028
	(0.065)	(0.069)	(0.036)	(0.036)	(0.081)	(0.090)	(0.046)	(0.035)
N (Total)	473	470	470	473	474	474	474	474
Mean of control group (Total)	15.681	16.115	-0.201	-0.225	15.757	16.150	-0.229	-0.165

Notes: This table decomposes the treatment effect on revenue collection (Columns 1, 5) into the effect on three components: changes in the Tax Base (Columns 2, 6); changes in the Non-Exemption Rate (Columns 3, 7); and changes in the Recovery Rate (Columns 4, 8). 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. Rows indicate the relevant margins of collection (total revenue, current year revenue, and collections against past arrears). Control variables include baseline tax base, non-exemption rate, and recovery rate. Outcome variables and controls in logs. Specification includes stratum fixed effects. The Information treatment is included in the control group. Number of observations and means of control group are reported for total collections (current and arrears are similar). 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 6: Impacts on Reassessments

Panel A

	(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

Panel B

				Components of GARV						
	(1) GARV	(2) Number of floors	(3) Last renovation was ≤ 2 years ago	(4) Land area (sq. feet)	(5) Total covered area (sq. feet)	(6) Main Road	(7) Tax Category	(8) Percent of property commercial	(9) Percent of property commercial and rented	(10) Tax Liability
Re-assess *	20674.778	0.002	-0.005	-271.548	869.811	-0.002	-0.220***	0.018	0.075**	4118.466
Treatment	(16481.084)	(0.050)	(0.020)	(746.256)	(769.903)	(0.048)	(0.084)	(0.037)	(0.029)	(3601.334)
Re-assess	24878.797***	0.078***	0.095***	334.908	-202.510	0.064***	0.204***	0.217***	0.176***	5517.176***
	(7786.877)	(0.026)	(0.011)	(514.958)	(376.675)	(0.024)	(0.041)	(0.019)	(0.015)	(1718.354)
N	15489	16352	16128	16352	16346	16352	15489	16226	16227	15489
Mean of control group in gen. pop. sample	35986.47	1.57	0.02	2703.99	2803.92	0.46	3.76	0.35	0.17	6483.80

Panel C

	(1) Approximate age of owner	(2) Owner's level of education	(3) Per-capita wages	(4) Predicted expenditure given assets	(5) Connected to Politician	(6) Connected to Politician/ Government/ Police
Re-assess *	-0.348	-0.523*	-821.749	111.044	0.021*	0.005
Treatment	(0.794)	(0.317)	(1078.070)	(213.404)	(0.012)	(0.027)
Re-assess	-0.656*	0.303*	13.126	-94.557	-0.013**	0.005
	(0.398)	(0.157)	(510.004)	(122.394)	(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	6291.64	0.05	0.36

Notes: This table examines whether the performance pay treatments affected the number of properties that were reassessed (Panel A), and how reassessed properties (Panel B) and property owners (Panel C) differed from the average property. The unit of observation is a circle, as defined at the time of the survey (Quarter 2 of FY 2012-2013). Panel A presents instrumental variables regressions, where treatment status is instrumented with randomization results. The sample consists of circles that were surveyed in the second phase of the survey (see Appendix B). Specification includes stratum fixed effects and controls for number of new and reassessed properties added in the pre-treatment (FY 2011) fiscal year. Panels B and C present instrumental variables regressions, where treatment status is instrumented with randomization results. 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. The characteristics in Panel B labelled Components of GARV are those that directly enter into the formula used to calculate GARV (see Appendix B for more information). Tax Category (Panel B, Column 7) is 7-tiered categorical variable with 7 being the most expensive tax bracket and 1 being the least expensive. Per-capita wages (Panel C, Column 3) is self-reported household expenditures divided by the total number of working household members. Predicted expenditure given assets (Panel C, Column 4) is the predicted value of a regression of household expenditure on series of dummy variables indicating various household assets. The Information treatment is included in the control group in all panels. Standard errors in all panels are clustered by the 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 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	-126.9	594.1*	.2021**	.0113
	(310.5)	(333)	(.0951)	(.0254)
N	9632	5993	4802	6050
Mean of control group	4919.067	1874.542	0.683	0.644
Panel B: Re-assessed and General Population Sample				
Re-assessed * Treatment	2248*	-557.4	-.1592*	-.0031
	(1311)	(367.1)	(.0934)	(.0221)
Re-assessed	3430***	-66.38	.0137	-.0191*
	(688.5)	(177.3)	(.0403)	(.0107)
N	13693	8207	6993	8268
Sample	Full	Phase 1	Phase 1	Phase 1
Mean of control group in gen. pop. sample	4713.484	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 8: 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	14182	14182
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 9: Mechanisms Beyond Price Effects

Panel A

	(1) Total	(2) Current	(3) Arrears
Information	0.069	0.071	0.226*
	(0.051)	(0.050)	(0.136)
N	482	482	479
Mean of control group	15.709	15.486	13.864

Panel B

	(1) Total	(2) Current	(3) Arrears
Revenue * Income Shock	0.0209	0.0499	-0.0481
	(0.104)	(0.111)	(0.218)
Revenue Plus * Income Shock	0.0445	0.0317	-0.250
	(0.169)	(0.143)	(0.362)
Income Shock	-0.0762	0.00846	0.119
	(0.0689)	(0.0461)	(0.0809)
N	478	478	478

Panel C

	(1) Total	(2) Current	(3) Arrears
Supervisory treatment	-0.063	-0.106	-0.035
	[0.329]	[0.232]	[0.848]
N	482	482	479
Mean of control group	15.924	15.679	14.176

Notes: This table presents results on mechanisms other than price effects that may be contributing to the observed impact of the performance pay schemes. In all panels, we use instrumental variables regressions, where treatment status is instrumented with randomization results. The unit of observations is a circle, as defined at the time of randomization. The outcome variable is log recovery as of the end of the second year of the study (FY 2012-2013). Columns separate recovery by total recovery (Column 1), current year recovery (Column 2), and collections against past arrears (Column 3). Panel A: This table re-estimates main revenue outcomes by subtreatment, with the Information treatment separated from the control group. Coefficients for the Revenue, Revenue Plus, and Flexible Bonus treatments not shown. Panel B: This table examines whether income effects contribute to the observed outcomes. The income Shock is calculated as the amount circle staff team would have earned under the scheme due to business-as-usual, plus their combined base salary, and is measured in logs (see Section for details). Specification follows Equation 6.4 of the main text. The coefficient for Flexible Bonus is not shown. Panel C: This table examines the impact of the supervisory treatments. Standard errors: For Panel A, we present 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. For Panel B, standard errors, in parentheses, are bootstrapped with 1000 iterations over the two-step estimation procedure (i.e. first estimating the income shock, and then estimating the model). The bootstrap sampling procedure is clustered by robust partition. For Panel C, randomization-inference p-values are shown in brackets, calculated over 1000 iterations. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 10: Cost-effectiveness of Incentives

	(1) Additional Revenue	(2) Cost of Incentives	(3) ROI
Panel A: Information in controls			
Any treatment	123,178,604	108,387,160	13.65
Revenue	49,862,425	37,349,784	33.50
Revenue Plus	40,180,674	35,549,342	13.03
Flexible Bonus	30,027,290	35,488,035	-15.39
Panel B: Information out of controls			
Any treatment	140,742,397	108,387,160	29.85
Revenue	56,107,778	37,349,784	50.22
Revenue Plus	45,522,473	35,549,342	28.05
Flexible Bonus	35,531,241	35,488,035	0.12

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.

Appendices

Table A.1: Balance

	Main Treatment					Information Treatment		Supervisory Treatment	
	Control	Treatment	Revenue	Revenue Plus	Flexible Bonus	Control	Treatment	Control	Treatment
Log Revenue	15.47	-0.037	0.024	-0.053	-0.055	15.46	0.050	15.65	-0.167
		(0.042)	(0.059)	(0.057)	(0.058)		(0.065)		(0.089)
		[0.347]	[0.683]	[0.375]	[0.378]		[0.451]		[0.156]
Log Recovery Rate	-0.333	-0.015	0.007	-0.002	-0.039	-0.330	-0.026	-0.366	0.006
		(0.024)	(0.035)	(0.034)	(0.040)		(0.037)		(0.041)
		[0.532]	[0.827]	[0.965]	[0.271]		[0.488]		[0.909]
Log Non-exemption Rate	-0.251	-0.024	-0.001	0.009	-0.059	-0.245	-0.035	-0.267	0.009
		(0.019)	(0.021)	(0.023)	(0.038)		(0.035)		(0.019)
		[0.200]	[0.964]	[0.721]	[0.0280]		[0.292]		[0.791]
Number of staff posted	2.564	0.055	0.038	0.056	0.088	2.576	-0.077	2.549	0.054
		(0.053)	(0.070)	(0.076)	(0.077)		(0.080)		(0.062)
		[0.300]	[0.600]	[0.494]	[0.280]		[0.343]		[0.607]
All positions filled	0.519	0.059	0.043	0.094	0.054	0.531	-0.072	0.538	-0.011
		(0.044)	(0.065)	(0.064)	(0.066)		(0.065)		(0.056)
		[0.188]	[0.506]	[0.146]	[0.414]		[0.267]		[0.908]
Log benchmark	15.44	-0.017	0.036	0.014	-0.073	15.44	-0.010	15.59	-0.114
		(0.044)	(0.062)	(0.059)	(0.064)		(0.074)		(0.095)
		[0.665]	[0.565]	[0.845]	[0.252]		[0.896]		[0.390]
FY 10-11 log growth rate	0.0280	0.003	0.017	-0.005	-0.005	0.0233	0.026	0.0259	-0.008
		(0.013)	(0.014)	(0.016)	(0.023)		(0.022)		(0.019)
		[0.824]	[0.378]	[0.813]	[0.813]		[0.217]		[0.788]
P-val, joint sig.		0.412	0.793	0.006	0.455		0.261		0.003
P-val, from RI		0.359	0.796	0.010	0.571		0.430		0.136

Notes: The table presents balance tests for the randomization into the different treatments. Columns labelled Control reflect control group means. Values in the treatment columns are the coefficients of a regression of the baseline value of the variable indicated in the row on a treatment dummy (or the set of subtreatments dummies), controlling for the relevant randomization strata. In the Main Treatment tests, the Information treatment is included in the controls. In the Information Treatment tests, the Information treatment group is compared against pure controls. The Supervisory Treatment test compares against its own control means (which are different from column 1 means). Robust standard errors in parentheses. Randomization inference (RI) based p-values in brackets. RI statistics are based on 1000 re-randomization iterations. * p<0.10, ** p<0.05, *** p<0.01. Stars reflect randomization inference based p-values.

Table A.2: Correlation of Satisfaction and Corruption Variables

	(1) Satisfaction	(2) Quality	(3) Bribes (normalized)	(4) Perception of Corruption
Panel A: Correlations with own response				
Satisfaction		.6808***	.0133	-.0407
		(.0193)	(.1084)	(.0295)
Quality	.7454***		-.261**	-.006
	(.0196)		(.1068)	(.0275)
Bribes (normalized)	3.7e-04	-.0066***		-.0066
	(.0029)	(.0018)		(.005)
Perception of Corruption	-.0191	-.0026	-.112	
	(.0141)	(.0117)	(.0748)	
N	5738	5738	5738	5738
Panel B: Correlations with others in circle				
Satisfaction	.8966***	.1041***	-.2206	.0014
	(.0149)	(.0147)	(.2047)	(.0045)
Quality	.0828***	.8598***	.1115	-.0019
	(.0153)	(.0167)	(.1347)	(.0048)
Bribes (normalized)	-.0025	5.6e-04	.7315***	-9.8e-04
	(.0019)	(.0012)	(.2412)	(.0011)
Perception of Corruption	4.0e-04	6.8e-05	-.0395	.9719***
	(.002)	(.0024)	(.05)	(.0016)
N	6400	6400	5738	6400
Mean of dependent variable	0.568	0.548	0.303	0.645

Notes: This table presents OLS regressions where we examine the relationship of our subjective non-monetary outcomes with each other (Panel A) and other respondents (Panel B). The unit of observation is a property. Panel A reports regressions of dependent variable on own response of independent variables. Panel B reports regressions of dependent variable on circle-level estimators of independent variables (where the own measure is excluded). Bribes are normalized by gross annual rental value (GARV) to be comparable within circle. Satisfaction, Quality, and Perception of Corruption are measured on 5-point Likert scaled, normalized to a [0,1] interval. 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 A.3: Inspectors' Knowledge of Treatments

Perceived Treatment	Revenue	Revenue Plus	Flexible Bonus	Information	Control	N
Actual Treatment						
Revenue	0.82	0.16	0.02	0.00	0.00	51
Revenue Plus	0.07	0.93	0.00	0.00	0.00	54
Flexible Bonus	0.04	0.05	0.86	0.00	0.05	56
Information	0.00	0.00	0.00	0.80	0.20	50
Control	0.03	0.02	0.00	0.01	0.94	159

Notes: This table provides a tabulation of inspector's understanding of their own treatment status. Rows list inspector's actual treatment status. Columns list inspector's perceived treatment status. Cells list the fraction of inspectors in the treatment given by the row who believe themselves to be in the treatment specified in the column (i.e. columns 1 through 5 should total to 1.00).

Table A.4: Impacts on Satisfaction with the Government

	(1) Quality of Electricity Dept.	(2) Quality of Water Dept.	(3) Satisfaction with Electricity Dept.	(4) Satisfaction with Water Dept.	(5) Likelihood of Picking up Note	(6) Indicated Preference for Incumbent Party
Panel A: Main Treatment						
Any Treatment	-0.006	-0.011	-0.003	0.003	-0.020	0.004
	(0.028)	(0.022)	(0.030)	(0.023)	(0.029)	(0.037)
Panel B: Subtreatments						
Revenue	-0.023	-0.031	-0.015	-0.039	-0.011	0.048
	(0.042)	(0.030)	(0.047)	(0.031)	(0.044)	(0.049)
Revenue Plus	0.057	0.039	0.050	0.074**	-0.005	-0.026
	(0.043)	(0.031)	(0.044)	(0.030)	(0.041)	(0.050)
Flexible Bonus	-0.051	-0.041	-0.042	-0.028	-0.041	-0.003
	(0.037)	(0.029)	(0.040)	(0.032)	(0.043)	(0.056)
N	4840	4840	4840	4840	4840	4840
Sample	Phase 1	Phase 1	Phase 1	Phase 1	Phase 1	Phase 1
Mean of control group	0.416	0.507	0.431	0.520	0.351	0.605
Rev. vs. Multitasking p.	0.575	0.343	0.707	0.065	0.801	0.226
Objective vs. Subjective p.	0.102	0.155	0.179	0.182	0.480	0.794
Equality of Schemes	0.091	0.070	0.206	0.004	0.771	0.433
Joint significance	0.175	0.124	0.363	0.012	0.821	0.630

Notes: This table examines the impact of the overall treatments (Panel A) and subtreatments (Panel B) on a wider range of non-monetary outcomes compared to those in Table 4. We present estimates from instrumental variables regressions, where treatment status is instrumented with randomization results. The unit of observation is a property. Quality and Satisfaction were measured on a 5 point Likert scale and re-scaled to a [0,1] interval. Likelihood of Picking up Note is the respondent's assessment (on a 5 point scale, also re-scaled to a [0,1] interval) of how likely a stranger would return them a Rs. 1000 they had accidentally dropped. Indicated Preference for Incumbent Party is a binary variable = 1 if the respondent indicates that any member of the household voted for the incumbent party at either the provincial or national level in the most recent elections. Sample restricted to circles surveyed in the first phase of the survey (see text for details). Specification includes strata fixed effects and controls for whether the property was surveyed using the short version of the survey. Information treatment 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 A.5: Impacts on Inspector Effort

	(1) Team Effort	(2) Hours/Day spent in field	(3) Hours/Day spent in office	(4) Total hours worked in typical day
Panel A: Main Treatment				
Any treatment	-.577	-.283**	.18	-.104
	(2.02)	(.141)	(.145)	(.0745)
Panel B: Subtreatments				
Revenue	-2.6	-.0745	.046	-.0285
	(3.78)	(.213)	(.215)	(.122)
Revenue Plus	2.39	-.286	.122	-.165**
	(2.69)	(.181)	(.19)	(.0732)
Flexible Bonus	-1.76	-.442**	.337*	-.105
	(2.83)	(.215)	(.203)	(.0959)
N	352	352	352	352
Mean of control group	90.6	5.46	2.68	8.14

Notes: This table examines the impact of performance pay on self-reported inspector effort. We use instrumental variables regressions, where treatment status is instrumented with randomization results. The unit of observation is a circle, as defined at the end of Year 2. Team Effort is assessed on a 100 point scale. The Information treatment is included in the controls. 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. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

A Generalizations of Model

In the model presented in Section 3, we assumed for simplicity that the costs of reducing tax liabilities were linear in the amount reduced (i.e. the cost for taxpayers is $\alpha(\tau^* - \tau)$ and for inspectors is $\beta(\tau^* - \tau)$. Here we show two generalizations of this setup. In Section A.1 we generalize this cost function. In Section A.2 we derive a related model where the costs are in terms of bribes paid, rather than reductions in tax liability. The general conclusions are unchanged from the simple version in the text.

A.1 Generalizing the Cost Function

Suppose that the taxpayer's cost of accepting a reduced tax liability is $\alpha f(\tau^* - \tau)$ and the tax inspector's cost of giving a reduced tax liability is $\beta g(\tau^* - \tau)$. We make only two restrictions on f and g. First, we assume that the marginal cost of reducing tax liability is always positive, and second, we assume that it is always weakly increasing. Specifically, we assume that $f'(0) > 0$ and that $f'' \geq 0$, and analogously for g.

With these assumptions, we can rewrite the surplus from agreement in (3.1) as

$$\tau^* - \tau - \alpha f(\tau^* - \tau) + r(\tau - \tau^*) - \beta g(\tau^* - \tau) \tag{A.1}$$

Taking the derivative with respect to τ, we obtain the following first order condition

$$1 - \alpha f'(\tau^* - \tau) - \beta g'(\tau^* - \tau) - r = 0 \tag{A.2}$$

One can see that the surplus from reducing tax liability will be positive (i.e. the derivative of the surplus with respect to τ will be negative) if and only if

$$1 - \alpha f'(\tau^* - \tau) - \beta g'(\tau^* - \tau) > r \tag{A.3}$$

Several observations are worth making about the condition in (A.3). First, as before, with $r = \alpha = \beta = 0$, i.e., with no incentives and no costs of reducing liability, collusion will always take place. Second, given the conditions on f and g, which ensure that f' and g' are always positive, equation (A.3) shows that there will be a tradeoff such that as r increases, the range of (α, β) such that any collusive equilibrium can take place shrinks. This can be shown by checking whether, starting from no collusion ($\tau = \tau^*$), there is scope for collusion, since if there is no scope for collusion at $\tau = \tau^*$ there will not be at any lower level of τ, since $f'' \geq 0$ and $g'' \geq 0$. At $\tau = \tau^*$, equation (A.3) reduces to

$$1 - \alpha f'(0) - \beta g'(0) > r \tag{A.4}$$

Since both $f'(0) > 0$ and $g'(0) > 0$, this shows that as r increases, the range of α and β where collusion is possible will fall.

It is also worth noting that with general cost functions, it is no longer the case that, conditional on collusion, one always sets $\tau* = 0$. Instead, the equilibrium amount of tax reduction τ^* is given by setting equation (A.1) equal to 0, which could be at an interior level of τ if $f'' > 0$ or $g' > 0$.

In this case, it is no longer obvious that the response to an increase in r will be an increase in bribes conditional on staying in the collusive equilibrium, because now it is possible that, instead, τ might increase (i.e. evasion would decrease). To see this, applying the implicit function theorem

to (A.2) yields

$$\frac{\partial \tau}{\partial r} = \frac{1}{\alpha f'' + \beta g''}$$

Thus, in general, so long as $f'' > 0$ or $g'' > 0$, taxes paid increase with r even for those who continue in the collusive equilibrium, since the marginal costs of collusion have increased.

What about bribes paid conditional on remaining in the collusive equilibrium? Recall that bribes b are equal to the foregone outside option of the inspector, $r(\tau^* - \tau)$, plus a share γ of the surplus. There are now two offsetting effects with an increase in r: the direct effect that the outside option increases will tend to increase bribes (i.e. $r(\tau^* - \tau)$ increases); on the other hand, if τ increases precipitously, the surplus will decrease and this could decrease bribes. Which of these effects dominates depends on the steepness of f'' and g'' – if f'' or g'' is sufficiently high bribes may actually decrease, whereas if they are quite low (as in the linear case in the text where $f'' = g'' = 0$) bribes will increase.

For example, suppose that $g(\tau^* - \tau) = 0$ and $f(\tau^* - \tau) = (\tau^* - \tau)^2 + (\tau^* - \tau)$, which satisfies the conditions that $f'(0) > 0$ and $f'' > 0$, and $f(0) = 0$. To find the optimum level of τ, we substitute f into equation (A.1), take the derivative with respect to τ, and set it equal to 0. This yields that $\tau^* - \tau = \frac{1-\alpha-r}{2\alpha}$. Note that conditional on evasion taking place, which occurs whenever $1-\alpha > r$, the extent of tax evasion in this setup is decreasing in r and α. Next, note that bribes are equal to $\gamma(surplus) + r(\tau^* - \tau)$, which in this case is equal to $\gamma\left[(1-r)\frac{1-\alpha-r}{2\alpha} - \alpha f\left(\frac{1-\alpha-r}{2\alpha}\right)\right] + r\frac{1-\alpha-r}{2\alpha}$. Simplifying yields $b = \frac{1}{4\alpha}(1-r-\alpha)(2r + \gamma(1-r-\alpha))$. Taking derivatives with respect to r yields $\frac{\partial b}{\partial r} = \frac{1}{2\alpha}(1 - 2r - \alpha - \gamma(1-r-\alpha)))$. Even in the range where bribes occur (i.e. where $(1-r-\alpha) > 0$), the sign of $\frac{\partial b}{\partial r}$ is ambiguous. Thus, in a more general model, one can obtain the predictions that government tax revenues τ unambiguously increase with r while the impact on bribes is ambiguous, even when all taxpayers stay within the collusive equilibrium.

A.2 Specifying Costs Based on Bribes Rather than Evasion

In the model in the text, the cost to taxpayers was in terms of tax evasion, i.e. in terms of $(\tau^* - \tau)$. An alternative specification of the model would be to specify the costs in terms of the bribes paid. i.e. in terms of b. This is similar to the model in the text, but slightly more cumbersome to work with, since now when solving the Nash bargaining one needs to take into account the fact that the transfers between the parties, b, are differentially costly for both sides. This means that the model can no longer be solved by computing the surplus and assigning a share γ of it to the tax inspector plus his outside option, but must be solved directly by maximizing the product of the surpluses.

Specifically, in this model, we suppose that the utility of the taxpayer is $-\tau - b - \alpha b$ and the utility of the tax inspector is $r\tau + b - \beta b$. Note the difference between this setup and the utility specified in Section 3 – now there is no utility penalty from tax evasion per se, but instead in this model bribes are less valuable than money, reflecting perhaps the possibility of being caught or the need to launder the bribes to evade detection. To simplify the algebra, we assume equal bargaining weights (i.e. $\gamma = \frac{1}{2}$). Since this model is linear, as in the model in Section 3, conditional on collusion one will always set $\tau = 0$ and fully evade.

To solve the problem, the taxpayer and the tax inspector maximize the product of their joint surplus, i.e. solve

$$\max_b \{r\tau + (1-\beta)b\}\{-\tau - (1-\alpha)b\} \tag{A.5}$$

subject to the constraint that $0 \leq b$ and $0 \leq \tau \leq \tau^*$. This yields the solution that collusion takes place if and only if $\alpha < \frac{1-\beta}{r} - 1$. Conditional on collusion taking place, $\tau = 0$ and $b = \left(\frac{1}{2(1+\alpha)} + \frac{r}{2(1-\beta)}\right)\tau^*$. Note that the same qualitative results from the model in Section 3 apply here as well: the range of parameters at which collusion takes places is decreasing in r, α, and β, and conditional on collusion taking place, bribes are increasing in r.

This model easily accommodates an effort margin as well. Suppose that, prior to bargaining with the taxpayer, if the tax inspector exerts 0 effort, he learns that the property is valued at an amount τ_1. If he exerts effort e, with probability e he discovers that the true property valuation is $\tau_2 > \tau_1$. His cost of effort is $\frac{1}{2}ce^2$. Once the valuation τ_1 or τ_2 is revealed, this is common knowledge between taxpayer and tax inspector, and they jointly solve the bargaining game in (A.5). In this context, effort is also increasing in the incentive rate r, regardless of whether the equilibrium is collusion, full payment of taxes, or a combination.

B Data

We use two main sources of data for analysis. Section B.1 describes the administrative data we use as our main measures of tax performance, and Section B.2 describes the survey data we conducted to obtain measures of accuracy of tax assessment, customer satisfaction, and corruption.

B.1 Administrative Data

Our primary data is quarterly administrative data on tax collections. Each quarter, as part of their normal reporting requirements, circle inspectors report their revenue collected during the fiscal year cumulatively through the end of the quarter, which they compile from tax paid receipts retrieved from the national bank. In addition, they also report their total assessed tax base before exemptions are granted (known as "gross demand") and after exemptions have been granted (known as "net demand"). These records are compiled separately for current year taxes and arrears. We digitized these quarterly reports for each of the approximately 500 tax circles in Punjab for a total of 6 years (4 years prior to the project beginning and the 2 years the project was in place).

Given the importance of this data in determining payments and measuring impact, it is important to validate its accuracy. Since reported tax receipts have to match actual money deposited in the bank and ultimately, received by the provincial Treasury department, the margin for discrepancies is low.[44] However, to ensure that the department's administrative data is correct circle-by-circle (since department's internal cross-checks are usually run at a higher level of aggregation), we instituted an additional re-verification program where we cross-checked the department's administrative records against the bank records. This entailed selecting a subset of circles (done identically in treatment and control areas), obtaining the individual records of payment received from the bank for each property, and manually tallying the sums from the thousands of properties in each circle to ensure that it matched the department total. Each circle had about an 18 percent chance of being verified by our cross-checks at some point during the two year experiment. We found virtually no systematic discrepancies between the administrative data we had received from the

[44] When a taxpayer pays his tax due at the local bank, in addition to a receipt that he retains as proof of payment, two additional receipts are generated and collected by the bank. One of these is returned to the tax department and the other is given to the Treasury. The latter's totals (at the district level) are then matched to both the department's aggregates and also to the actual amount transferred by the bank to the Treasury.

department and what we found in these independent verifications; the average difference between our independent verifications and what the circle had reported revealed under-reporting of -0.28%, or about zero.[45]

The administrative data also contains information on the identity of the inspector, which allow us to track if inspectors are relocated. We supplemented this by conducting a survey, each quarter, of the locations of all inspectors, constables, and clerks in the tax department.

B.2 Taxpayer/Property Survey Data

The second major data source is an independent property survey we conducted. This survey had three main purposes. First, it allowed us to obtain data on people's interactions with the tax department, both in terms of their overall perceptions of the quality of this interaction and on corruption.[46] Second, we obtained an independent assessment of the property's characteristics (e.g. land area, covered area, location), which we could use to construct an independent assessment of the property's valuation and compare to the department's official assessments.[47] Third, we obtained information about the owners and property characteristics which allow us to understand whether any observed impact of the schemes varied by the types of properties and owners.[48]

To do so, we surveyed approximately 16,000 properties. Properties were sampled in one of two ways. First, to obtain a general population sample of all properties (including those not necessarily on the tax rolls), we created GIS maps of the circle boundaries for all 482 circles, and used GIS software to randomly select 5 points within each circle. We then surveyed the property nearest that point, and selected 7 more properties nearby (chosen by walking left from the point and choosing every other property) of which an additional four were surveyed based on a randomization table. Once this was completed, we matched these properties to the property-level administrative data to obtain the corresponding administrative records for these properties. On average, 85% of properties we randomly sampled in the field could be matched to corresponding administrative records, suggesting quite high coverage of properties throughout the province. These properties, which we refer to as our "general population sample," represent a random sample of 25 properties per circle, or 12,000 total.[49]

[45] One complication is that circle boundaries are modified over time, as circles are merged and split to better reflect realities on the ground. In our data, out of the 482 circles present at the time of randomization, a total of 117 were affected by merges or splits throughout the 6 year period covered by our administrative data. To maintain consistency, we reconstruct the data at the level of the 482 circles present at the time of randomization. For those circles that merged prior to randomization, or split post randomization, one can simply add the two split circles together to obtain correct values for circles with randomization-era borders. For circles that split prior to randomization, or merged after randomization, we use the ratio of current year tax base net of exemptions (called "net demand" by the department) among the new and old circles reported in the quarters immediately before and after the split/merge to apportion the new circles to the randomization-era circles. The main results are qualitatively similar if we instead simply restrict analysis to the 365 circles that were unaffected (see Online Appendix Table 3-B).

[46] The quality of interaction questions are solicited on a traditional Likert scale. For corruption, given that respondents are often not comfortable revealing their own bribe payments, we ask about the incidence of corruption and amount of bribe that would need to be paid for a property similar to theirs.

[47] Assessing areas is relatively straightforward since most properties in Punjab use standard lot sizes. To calibrate this, surveyors practiced assessing the size of various sizes of properties in training so that they could reliably estimate property sizes.

[48] In cases when the property was rented, we were not always able to obtain information about the owner. Although renters form 25% of our sample, Online Appendix Tables 4-H and 7-H show that the results are qualitatively similar if we drop rented properties.

[49] For budgetary reasons, one-fifth of the surveys in a circle were conducted using a shortened version of the

Second, since we were particularly interested in the properties whose tax valuation had changed, we also sampled properties directly off the separate tax lists that are maintained for newly assessed or re-assessed properties to ensure we had sufficient representation of these properties in our sample. Specifically, we randomly selected 10 properties in each circle from those that had been re-assessed during FY11-12 and FY12-13, and then located these properties in the field and surveyed them. We denote this sample of over 4,300 properties as the "re-assessed" sample.

The survey was conducted at the end of the experiment. For logistical reasons it was split into two phases. The first phase was conducted during June and July 2013 (with a few properties finished in August and September), and covered approximately half the circles (randomly selected). The second phase was from October 2013 to January 2014, and covered the remaining half of the circles. For subjective measures (e.g. bribes, customer satisfaction), we focus on the results from the first wave of the survey, when respondents would surely be answering for the correct time period when the treatments were in effect. For objective measures (e.g. accuracy of assessment, property characteristics), we use both survey waves.

While the other measures are fairly self-explanatory, we should explain the assessment (in)accuracy measures. This is calculated by comparing the inspector's official assessment with the assessment we compute using our independent survey, normalized by the sum of the two measures, according to the following formula:

$$Inaccuracy = \frac{|GARV_{Inspector} - GARV_{Survey}|}{(GARV_{Inspector} + GARV_{Survey})} \tag{B.1}$$

This variable ranges from 0 to 1, with 1 indicating the greatest difference between the two metrics.[50] We also examine the tax gap, which is the same metric without the absolute value and which measures average amounts of over/under taxation:

$$TaxGap = \frac{GARV_{Inspector} - GARV_{Survey}}{(GARV_{Inspector} + GARV_{Survey})} \tag{B.2}$$

This measure ranges from 1 (complete over-taxation: inspector assesses the property positively whereas independent survey reveals 0 liability) to -1 (complete under-taxation: inspector assesses the property at 0 whereas independent survey reveals positive tax), with 0 indicating agreement.

questionnaire. The choice of which properties received short versus long survey was pre-determined as part of the sampling protocol and hence effectively randomized. When analyzing the survey data, we control for the format of the survey with a short survey dummy.

[50] Note that this measure is normalized by the sum of $(GARV_{Inspector} + GARV_{Survey})$. An alternative would just be to treat our measure, $GARV_{Survey}$, as the truth and normalize by that. However, if there is iid measurement error in each estimate, the average of the two will help smooth out the measurement error.