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

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

April 2014

Abstract

Despite the importance of performance-based payment systems, there is little rigorous evidence on their impact in a civil service. In tax, high-powered incentives for tax collectors could increase revenues, but come at a high political cost if collectors exert excessive pressure on taxpayers. We report the results of the first large-scale, randomized field experiment designed to investigate these issues. Working with the entire property tax department in Punjab, Pakistan, we experimentally allocated collectors in the entire 482 property tax units into one of three performance-pay schemes or a control group. We find that incentivized units experience an average of 9 percentage points higher revenue growth than controls, and the effects persist over the two years of the experiment. Yet these units do not report greater taxpayer discontent. The performance pay scheme that rewarded purely on revenue collection does better, increasing revenue growth by 14 percentage points, with relatively little penalty for customer satisfaction and assessment accuracy compared to the two other schemes that explicitly rewarded these dimensions. Digging deeper into our findings reveals an interesting heterogeneity: most taxpayers in incentivized areas do not get reassessed or pay higher taxes, but do report higher level of bribes and instances of corruption. In contrast, conditional on their property being reassessed, taxpayers in incentivized areas pay higher tax and report less corruption. These results are consistent with a simple model of collusion between collector and taxpayer, in which taxpayers in incentivized areas either have to pay higher bribes to avoid being reassessed, or pay substantially higher taxes if the bargaining process breaks down. Our results suggest that while incentives for tax collectors can substantially boost revenue collected and lead to little overall discontent, they do empower the tax collector in a manner that can both lead to increased revenue and rents at the expense of the taxpayer.

PRELIMINARY AND INCOMPLETE.

*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, and Adeel Shafqat for outstanding research assistance in Lahore. We thank all the Secretaries, Director Generals, the two Project Directors from the Punjab Department of Excise and Taxation, the Punjab Finance, Planning and Development departments and the Chief Secretary and Chief Minister's offices for their support over the many years of this project. Financial support for the evaluation came from 3ie, the IGC, and the NSF (under grant SES-1124134), and financial support for the incentive payments described here came from the Government of the Punjab, Pakistan. The views expressed here are those of the authors and do not necessarily reflect those of the many individuals or organizations acknowledged here.

1 Introduction

Tax systems throughout the developing world collect substantially less revenue as a share of GDP than their counterparts in developed countries (Gordon and Li, 2009). While there are many sources of these differences, an important one may be in the role played by the tax officials who assess, audit, and enforce taxes. Due to the largely manual nature of tax administration in many developing countries and the comparative lack of third-party reporting, these officials play an important role in determining tax liabilities and enforcement. But, given 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. But these regimes were often unpopular, as tax farmers used aggressive tactics to collect taxes. In fact, unhappiness over the aggressive techniques and substantial rents accruing to tax farmers was a contributor to the French Revolution, and tax farmers were among those guillotined in 1794. US states similarly experimented with highly incentivized "tax ferrets" to collect property taxes in the 19th century, but again this fell out of favor. More generally, 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), there has been relatively little appetite for revisiting this debate, perhaps because the perceived benefits are outweighed by the costs. 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, Pomeranz and Singhal, 2014), this tradeoff is not as obvious. Not surprisingly some countries such as Brazil, Pakistan, and others have begun to reconsider incentives for tax staff. Yet there is little rigorous evidence of the tradeoffs developing country governments face between increased revenue from incentives and the political costs that may occur as a result of a disgruntled citizenry.

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, consisting 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 up to two full fiscal years. The incentives were large: the 3 tax personnel in each circle were collectively given an average of 30 percent of all tax revenues they

collected above a historically-predicted benchmark.[1] Many personnel in treated areas were able to double their base salaries through these incentives .

Given the historical concerns about potential negative impacts of high-powered incentives on taxpayer satisfaction and fairness, the three schemes varied in both the extent to which they also based performance pay on non-revenue outcomes (as a means of potentially mitigating political cost on taxpayers) and the extent to which they allowed for more 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. The "Revenue Plus" scheme provides incentives exactly as in the Revenue scheme, but makes adjustments equal to about +/- three-fourths of base salary based on whether the circle falls 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. Explicitly providing these multi-dimensional incentives might mitigate some of the political costs that were the downfall of tax farming historically. The third scheme, "Flexible bonus," took this a step further by both rewarding collectors for a much wider set of pre-specified (by the tax department) criteria, 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, then there will be two different effects: some taxpayers will continue in the collusive, low-tax equilibrium but with 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. Performance pay thus could have heterogeneous effects on tax revenue and bribes among taxpayers.

We evaluate the impact of the schemes using multiple sources of data. To evaluate the impact on tax revenues, we obtained quarterly administrative data on tax revenue collected for each circle, which we verified by conducting random spot-checks against the tax department's bank records. To evaluate the impact on other outcomes, such as corruption and tax accuracy, we conducted a survey of over 16,000 properties throughout the province, in which our surveyors asked property owners about their perception of corruption in and sense of satisfaction with the tax department. The surveyors also directly observed the characteristics of each property that are used in the tax

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

calculation, and then manually matched each property to the tax rolls to obtain the corresponding tax records for each property. Since tax assessment is determined formulaically from these property characteristics, with no subjectivity on the part of the inspector, we can determine the accuracy of assessments by comparing our measurements to those on the official tax rolls.

We find that, on average across the three schemes, performance pay led to an increase in tax revenue of about 8 percent based on the administrative data. We show that this came predominantly through an increase in the tax base (i.e. the total assessed values of properties), through both increased re-valuations of existing properties and the addition of new properties to the tax rolls, with some some additional impact on recovery rates of outstanding arrears in the first year. On average, we find no impact of the schemes on non-revenue/political costs. Specifically, the increased revenue generated as a result of the schemes is not accompanied by a decline in the average 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 change on average in the accuracy of tax assessments. Thus, on average, we find that the incentives increase revenue with little obvious downside in terms of legitimacy 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 detectable differences on non-revenue outcomes. Specifically, the Revenue scheme, which provided incentives purely based on revenue collected, increased current-year revenues by 17 percent. The Revenue plus scheme achieved only 9.5 percent and the Flexible Bonus scheme achieved only a statistically insignificant 4.2 percent increase. The Flexible Bonus scheme did not do better on any dimension we can measure in our data. Thus, adding multiple dimensions to performance pay diluted the impact on revenue to the extent that it was no longer as cost-effective for the government, and the flexible scheme that allowed for subjective performance assessments was almost entirely ineffective.

Our setting also allows us to dig deeper by examining how being under a performance pay regime impacts taxpayers. We find heterogeneity in response among taxpayers that is consistent with there being collusion between tax inspectors and taxpayers. Specifically, we find that for most properties in performance pay tax circles, taxpayers were not reassessed and yet they reported that the typical bribe paid to property tax officers for properties similar to theirs was Rs. 600 (about US $6) higher than in control areas, about a one-third increase compared to the control group. They also reported bribe payments were more frequent, but there was no change in their tax payments.

Yet, for the small number of properties whose tax valuation was formally changed (either newly assessed or re-assessed), these taxpayers report paying higher taxes, and do not experience the higher bribes that other properties in performance pay circles experience. Moreover, while the data suggests under-taxation for typical properties, this is undone for these properties, who appear on average to be taxed accurately. Moreover, there is an increase in the number of these new 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 tax bribes to compensate them for their forgone (performance) pay. But, given the higher bribe required to obtain a tax reduction, some taxpayers may instead prefer to switch from the collusive (low-tax, high bribe) to non-collusive (high-tax, low bribe) equilibrium.

These results suggest that the increase in tax revenue is driven by a relatively small number of properties that are (correctly) re-assessed and who switch from the collusive to non-collusive equilibrium, 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 re-assessed properties have taxable value that is about 70 percent higher than the typical (non-reassessed) property. In treatment areas, the reassessed properties are even more valuable than re-assessed properties elsewhere, by another 56 percent (p-value 0.001). Re-assessed properties in general are also more likely to be commercial properties, which are taxed at a higher rate. And, there is suggestive evidence that while those 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 that seeks to maximize tax revenues subject to political constraints, the evidence presented here suggests that the Revenue scheme is the most effective: it raised current-year revenue by 17 percent, which implies a substantially positive return-on-investment, and it did not appreciably reduce satisfaction with the tax department compared to controls.[2] While the Revenue Plus scheme did slightly better on satisfaction than the Revenue scheme, the cost in forgone revenues was such that the scheme was no as positive for the government.

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 tax payers. Section 4 outlines the experimental design, Section 5 describes the data and empirical approach, and Section 6 presents the results. Section 7 concludes.

[2]The relatively small impact on satisfaction may be because the revenue increases come from a very small number of properties, so there is relatively little average effect. Moreover, the model suggests that those who switch from the low-tax collusive regime to the high-tax non-collusive regime are precisely those who have a higher disutility from paying bribes. This switch may cause a second-order utility loss to taxpayers (since they are close to indifferent between the high-bribe, low tax state and the low-bribe, high tax state) but a first-order increase in revenue for the government.

2 SettingProperty 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 such as a narrow tax base, inadequate (in reflecting market value) and infrequently updated valuation tables, generous exemptions, tax evasion and corruption, distrust of public institutions, and administrative weaknesses (Bahl, Wallace and Cyan 2008,GHK 2009, World Bank 2006, World Bank 2009).

The 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 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 it 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 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 a significant impact these margins have on tax assessment and also given that it is less easy to verify whether a residential property is being rented or not 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 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).

2.1 Property tax administration

The primary unit of tax collection is the "tax circle," a predefined geographically contiguous area that covers anywhere from a two to ten thousand unique properties/taxpayers. The substantial variation in circle both in terms of property numbers and total assessed value is due to differential usage and growth of urban areas and a slow-moving circle "boundary redrawing" process. Within each circle there are three designated tax officers who work together as a team: an "inspector," who 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. Together, they maintain a record of all properties, apply valuation tables to individual properties, and determine what 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. The inspector is the head of this three-person team. 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.

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. Moreover, for a typical tax collector there are few avenues for vertical mobility through promotions. While there are desired circle postings (larger tax circles and those in more urban areas), transfers are often more politically based rather than merit based (Piracha and Moore, 2013).

However, unlike other departments such as education and health, the tax department does have explicit financial targets it has to meet each year. While these are mostly based on a given year's budgetary requirements or a simpler year-on-year increase (without any way to account for actual urban growth rates), these targets are passed down to each (of several) progressively smaller administrative tax unit all the way down to the smallest unit, the tax circle. Since each higher unit has corresponding tax officials, the normal mode of collecting against these targets is through each administrative level pressuring and cajoling lower levels to increase collections. Ultimately since the three-person tax circle team is the one with primary access to information about the taxpayers, these tactics would need to impact their behavior to have an impact. But, with limited reward mechanisms, threats of transfers are the primarily tool available, but since these are in turn largely based on political pressure these are not always credible to carry out.

The system thus 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/inspection mechanisms. The property database is manually recorded on physical registers, and does not automatically include new properties or property updates. Since there is no automatic link between building permits (when issued) and tax rolls, there is no way for the government to learn about these changes due to construction or changes in property use unless they are detected

and officially reported by 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 mismeasured. Finally, the manual system of billing and collection, 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, verification of how the tax officials at circle level use their discretion is extremely difficult.

Should they want to increase total revenue collected (say, in response to incentives), tax officials can do so in a variety of ways. Collectors likely have substantial private information regarding a property's true tax liability that they may choose not to reveal to the authorities. Thus schemes that reward on revenue collected may increase tax collected by incentivizing tax collectors to reveal such information. Extreme forms 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 (and higher) tax valuation of a property or denying (incorrectly provided) exemptions. The same margins can also respond to inspector effort in obtaining more accurate information on the new construction or changed status of properties or simply by visiting non-paying households to encourage them to pay, through a combination of social pressure and the threat of legal sanctions (such as summons by the tax office or, ultimately, sealing the property so it cannot be accessed until tax bills are paid).

3 Conceptual Framework

Before discussing the experimental design in detail, it is worth considering briefly how incentives for tax officials would affect both revenues collected by the government and bribes paid from taxpayers to inspectors.

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

Both taxpayer and tax inspector face some costs from colluding to report $\tau < \tau^*$. For the moment, we assume the costs for both are linearly increasing in $\tau^* - \tau$. Linearity is not crucial for the main results here; instead, for the qualitative patterns we discuss, all we need is that the marginal costs of collusion to both parties weakly positively increasing in $\tau^* - \tau$, and are for at least

one party strictly positive at $\tau^* = \tau$. For simplicity of exposition, we assume that 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 see 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

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

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 $\tau = \tau^*$.

Let γ be the bargaining weight of the taxpayer. 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 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 (i.e. the cost to the taxpayer of accepting a reduced tax liability is $\alpha(\tau^* - \tau)$). The linearity of the model gives leads to corner solutions – specifically, 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 generalize the model to allow for

more general cost functions, imposing the restrictions 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 general results shown here continue to apply: as r increases, tax revenue continues to increase, but the impact on bribe payments is ambiguous. In this 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.

It is also worth noting that in this model, the costs to the taxpayer and inspector are in terms of deviations from the true tax liability, i.e. $\alpha(\tau^* - \tau)$. An alternative specification would have been to specify that the costs are in terms of bribes paid and received, i.e. αb, 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 the same qualitative results as the in the model derived here obtain if we specify the costs in terms of bribes rather than in terms of tax evasion.

4 Experimental Design

To examine the role of performance incentives, tax circles were randomly allocated to one of three incentive programs, or a control group. 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 in the second year). In addition, in the second year of the experiment, two additional treatments were added: a performance pay scheme for supervisory personnel, and an "information-only" scheme that served as a placebo.

We begin by describing the three performance pay schemes each in turn. We then describe the additional treatments (supervisor and information-only scheme), and then conclude with a brief discussion of how well tax officials understood the schemes to which they had been assigned and the discuss the degree to which they believed the schemes would be implemented as described.

Revenue-based

This performance pay group rewards tax circle staff (inspectors, constables, and clerks) based on the revenue they collect above a predicted benchmark. Each circle is assigned one of three bonus

rates i.e. either 40%, 30% or 20%. The differential bonus rate take into account the differences in the size of circles; i.e. to promote equity (so that staff who happen to be in large circles would not receive vastly large bonus payments than staff in smaller circles)staff in larger circles are compensated at a lower rate than smaller circles where it perceived to be more difficult to raise a given amount of revenue.[3] The benchmark for each circle is generated using historical revenue data for that specific circle.

Specifically, each inspector continues 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.[4] It is important to note that this scheme treats increased collections due to expansion of the tax base (new properties) or increased collection on the current base (higher recovery rates) symmetrically. Benchmarks are 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.[5]Since most inspectors are rotated to new circles every three-to-four 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, since by the time higher revenue collected started 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 is divided 40%-30%-30% among inspector, constable, 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 faces a roughly 10% individual marginal incentive . For the modal circle, increasing revenue by 15% would result in a bonus payment to each staff member that could almost double their base

[3]Assigning a higher percentage to smaller circles is not necessarily optimal, but the decision was made by the tax department primarily for reasons of equity, rather than reasons of optimal incentive design.

[4]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 $Incentive_c = \alpha_c \max(CurrentYearRevenue_c - Benchmark_c, 0) + \alpha_c \max(ArrearsRevenue_c - ArrearsBenchmark_c, 0)$ 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.

[5]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 percent 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. 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%.

salary. 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.

Revenue Plus

The second performance pay scheme, Revenue Plus, is similar to the Revenue-based scheme, but includes explicit multi-dimensional incentives to help address the multi-tasking 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 political costs of taxes (i.e. measured by how people well feel they are being treated by the tax department) and whether taxes are being applied equitably.

To address these concerns, in this scheme, the tax officials receive the exact same incentives on revenue as in the Revenue scheme, but with adjustments based on taxpayer satisfaction and accuracy of tax assessments, assessed from an independent survey of properties in the circle (described in Section 5.2 below). Accuracy is measured as the 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.[6] Taxpayer satisfaction was measured based on two survey questions about the quality of interactions with the tax department and the results of ones interactions with the tax department.[7] Circles in the scheme were then ranked based on their accuracy and satisfaction measures and divided into three equal-sized groups. Circle staff get paid as in revenue treatment, but the top group gets an additional bonus equal to 0.75 times their base salary, and the bottom group lose 0.75 times their average base salary.[8] Note that 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) the average payments would be identical between the Revenue and Revenue Plus schemes.

Flexible Bonus

The third scheme aims to mimic the way bonuses work in the private sector for many complex jobs, such as those in Wall Street firms: managers distribute a fixed bonus pool to talented em-

[6] 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 subscore was based on the stength of this correlation.

[7] 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 scale from 1 to 5.

[8] Specifically, 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.

ployees based on all factors they observe. 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, so could adjust for these factors in ways that would be difficult in an objective, ex-ante specified formulaic incentive system. While such subjective assessments can potentially better match the complexities of real jobs, they can be less effective than formulaic systems if workers do not trust the managers to implement them properly, if managers play favorites, or if managers and workers disagree about the subjective component of performance (Baker, Gibbons and Murphy, 1994; Prendergast and Topel, 1996; Prendergast, 1999; Bentley, 2003)).

In this treatment, staff are again divided into three groups (just as in the Revenue Plus scheme), but rather than have pay determined by an ex-ante specified formula, they are divided by their performance as ranked by the departmental "Performance Evaluation Committee." Specifically, everyone in the treatment provisionally earns a base salary supplement roughly equal to their average salary (i.e. each inspector is entitled to an honorarium of Rs. 22,000/month, and each constable and clerk is entitled to an honorarium of Rs. 16,500/month).[9] At the end of the year, adjustments are made just as in the Revenue Plus scheme: the top third of circles gets an additional bonus equal to 0.75 times their base salary, and the bottom group lose 0.75 times their average base salary.[10] The Performance Evaluation Committee is allowed to use any criteria it chooses, so long as it can document as 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 two differences between Flexible Bonus and Revenue Plus are that the objective revenue-based formula is replaced by the fixed 1.5 times base salary average amount, and that the grouping is made by the Performance Evaluation Committee as it sees 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 Performance Evaluation Committee full flexibility in using subjective criteria, they in fact created a formula for ranking circles, using the following features: increase in revenue collected (40 percent), increase in tax assessments (25 percent), subjective director's rating (10 percent), customer satisfaction (10 percent), and accuracy of assessment (15 percent.) The categories to be used were 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) are tax base increases and the subjective director's assessment. On net, the correlation between the PEC ranking

[9] 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 figure was adjusted in the second year in order to better ensure that the schemes generated equal average honorariums.

[10] Specifically, inspectors in the top group received an extra Rp. 15,000 per month, and constables and clerks received an extra Rp. 11,500 per month; those in the bottom group lost an equivalent amount. In practice, since we it was not feasible to actually take money back once paid, circle staff were only paid 50% of the honorarium earned each quarter 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 below.

and the ranking of payments under the Revenue Plus scheme was 0.555 in Year 1 and 0.310 for Year 2. Additional treatments

In addition to the main circle-level performance pay treatments, which were in place for two fiscal years, we also introduced two additional treatments in the second year of the program (FY12-13). First, we introduced an "information-only treatment." This treatment was intended to see if there was any effect simply from providing the same information that was given to participants in the performance pay schemes (in particular, the project gave them quarterly reports on their collections above their historically-predicted benchmarks), but with no payments attached . While this information was compiled from what the circle staff provided to us and therefore it was not new information to them (though it may have been presented in a more systematic format), there could have been perceived monitoring coming from being selected under one of the treatments (and meeting each quarter). This scheme therefore differentiates between the payments per se, and any additional monitoring that may have been perceived to be associated with the treatments.

Second, a supervisor's performance pay scheme (for the next two tax administrative levels) 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 AETOs. Randomization was done at the level of the ETO, with 26 treatments and 25 controls. All AETOs working under selected ETOs were selected. Payments were identical to that of inspectors in the revenue scheme, except that they were paid based on the average of all circles under their supervision, and they received a 50 percent share of all imputed bonus payments, rather than the 40 percent share received by inspectors. 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.

Knowledge 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 the year, post-training quizzes and refresher trainings. These were carried out for the three performance-pay schemes, the supervisory scheme and the information treatment . 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) confirms that inspectors could accurately identify whether they would receive payments, and which scheme they were in. Additional details can be found in Appendix B.

Credibility of schemes and payment process

In addition to ensuring that staff understood the scheme they were in, since performance-based pay had never been introduced in the tax or for that matter any other Punjab department before, ensuring credibility of the schemes – i.e. that inspectors believed they would be paid as promised – was an important issue.

We did this through several means. First, in the year preceding the first year of the project, we conducted a pilot project with only 11 tax circles selected for incentive schemes. The fact that these circles were in fact paid each quarter according to the promised schemes added significant credibility that the same would hold when the larger project was rolled out. Second, there were accompanying concerns regarding the fairness of the whole project. Since these schemes could only increase one's pay, there were highly popular and most staff wanted to opt-in. In order to ensure that there was no perceived bias in this process of selection, the randomization was carried out through a public lottery, with a large number of taxation department present. Third, at each stage of the process, formal approvals were received from all relevant government departments, including the Chief Minister (the equivalent of the provincial governor), the Tax department, the Finance department, and the Planning and Development department, were received and communicated to all parties concerned.

Finally, the payment process was designed in a way to further establish credibility. Although the final payments each year were determined only based on end-of-year totals, staff were paid each quarter based on their cumulative earnings under the scheme through that quarter, with correspondingly quarterly benchmarks computed in the same way as the final annual benchmark.[11] Payments were carried out separately in each division with checks handed out to every staff member along with a detailed calculation of the amount paid so that the staff could verify that the amount was indeed correct.

4.2 Randomization and balance check

As noted above, randomization took place through public lotteries. In the first lottery, circles were selected to participate in the program or not. After being randomly selected to participate, informed consent was solicited from inspectors, constables and clerks, who were told about the three possible incentive schemes while making clear that a second lottery would determine which scheme they would be assigned to.[12] Once consent was obtained, a second lottery was held to assign

[11] To do this, we computed benchmarks for each quarter in the same manner as we computed annual benchmarks, and made payments based on the cumulative amount of revenue collected through that quarter compared to the analogously computed benchmark. Only half of cumulative earnings were paid out to mitigate the possibiltiy that staff were overpaid in the event that the pace of collections slowed over the year. This process was clearly explained in advance to inspectors and formed part of the training.

[12] Given the crucial role played by the inspector in determining tax treatments, 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.

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 of the experiment, in the second year of the experiment both selection into the treatment, and assignment to sub treatments, were conducted in a single stage lottery.

Table 1 shows the overall evaluation 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 pay in year 2.[13] In addition, 70 circles were selected for the non-incentive "information-only treatment." Each of these lotteries was stratified within 19 strata, with strata based on the 11 administrative divisions of the province, subdivided by the three reward-rate bins that would be used if a circle was chosen for Revenue or Revenue Plus treatments in cases where there were 18 or more circles within a given division. Lotteries were conducted publicly by computer. The lotteries were held as close as possible to the start of the fiscal year on July 1.[14]

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. Column 2 shows the difference between all incentive treatments and controls. All comparisons are done on the original randomization (i.e. intent-to-treat) results. Standard errors are shown in parentheses, and randomized-inference based p-values are shown in brackets. None of the differences are statistically significant. The remaining columns show the 3 incentive sub-treatments (Revenue, Revenue Plus, and Flexible Bonus), information treatment, and supervisory treatments, all compared to their respective controls. Out of the 42 comparisons made (7 variables * 6 columns), only 1 is statistically significant at the 5 percent level (the non-exemption rate in flexible bonus), as one would expect by random chance. Looking scheme by scheme, the joint test for statistical balance shows statistical significance in one of schemes (Revenue Plus) compared to pure controls, even though none of the individual covariates are statistically significantly different. We show below that the main average effects of incentives do not seem to be driven by this one sub-treatment.

5 Data and Empirical Methodology

We use two main sources of data for analysis. Section 5.1 describes the administrative we use as our main measures of tax performance, and Section 5.2 describes the survey data we conducted

[13] Since this was not part of the policy to start with (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.

[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.

to obtain measures of accuracy of tax assessment, customer satisfaction, and corruption.

5.1 Administrative Data

The main source of the data for the project is quarterly administrative data on tax collections. Each quarter, as part of their normal reporting requirements, each circle inspector reports their revenue collected during the fiscal year cumulatively through the end of the quarter, which they compile from banks statements they have received. 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. The excise and tax department uses these records for its internal monitoring purposes, but only has them in aggregated form (that is, each assistant supervisor creates a similar summary report for the circles he oversees; each supervisor creates a summary report for each assistant supervisor he oversees, and so on). We obtained and digitized the original quarterly reports at the circle level 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). These circle-level administrative records form the core of our analysis.

One of the main advantages of performance pay in tax is that the outcome variable (tax revenues) is accurately observed, since it must correspond to money deposited in the bank and ultimately, received by the provincial Treasury department. 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 (the tax bill has three identical copies) that are 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.

Given these existing institutional checks against actual money received, discrepancies are quite rare. However, to ensure that the department's administrative data is correct at the circle-level as well (since the checks are usually run at a higher 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 a 17% chance of being verified at some point. In general, we found virtually no systematic discrepancies between the administrative data we had received from the 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.

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 106 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 376 circles that were unaffected, or aggregate all circle splits up.

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.

5.2 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. 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. Third, we could obtain information about the owners and property types which allow us to understand whether any observed impact of the schemes varied by the types of properties and owners.

To do so, we surveyed approximately 16,000 properties. Properties were sampled in one of two ways. First, to obtain a random 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. Survey enumerators were given the latitude and longitude coordinates of each point, and used a handheld GPS device to locate that point on the ground. They 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. These properties, which we refer to as our "random sample," represent a random sample of of 25 properties per circle.

Second, since we were particularly interested in properties whose tax valuation had changed as a result of our treatment, we sampled properties directly off the separate tax lists that are maintained

for newly assessed or re-assessed properties. Specifically, we randomly selected 10 properties in each circle among those that had been re-assessed during the FY11-12 and FY12-13 fiscal years, and then located these properties in the field and surveyed them. We denote this the "re-assessed" sample.[15]

The survey contained three main sets of questions. First, to compute an independent assessment of the property value, the surveyor noted all aspects of the property that feed into the property valuation, such as the land area, size of buildings, property use, etc.[16] Second, to assess corruption, we ask a series of questions about unofficial payments. Given that respondents are often not comfortable revealing their own bribe payments, the key question we ask is the typical bribe that would need to be paid for a property similar to yours.[17] Third, we collected a variety of characteristics about the owner, which we use to assess how the treatment differentially affects different types of households.[18]

The survey was conducted in two rounds, beginning at the end of the experiment. The first round was conducted primarily during June and July 2013 (with a few properties finished in August and September), and covered approximately half the circles (randomly selected). After a pause for the Eid holidays, the second phase of the survey was conducted 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.

5.3 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. 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 were transferred out to another circle and could not

[15] In practice, the sample sometimes includes fewer than 10 properties per circle. This is because we stratified sampling by semesters across the two fiscal year and some circles did not have any reassessed properties in a given sampling window. Taking into account this adjustment, the average number of properties per circle is still 9.5 across the reassessed sample.

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

[17] Specifically, we ask "On average, what is the total amount in unofficial payments that a property owner of a property similar to yours would make to the [...] department over an entire year for the purposes above (and anything else)?" where the listed purposes are to reduce a bill, settle a dispute, obtain speedier service, obtain a permission to expand property, or change the use of a property."

[18] 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 **??** and **??** show that the results are qualitatively similar if we drop rented properties.

participate.

Specifically, our primary specification for assessing circle-level outcomes using the administrative data is

$$\ln Y_{cst} = \alpha_s + \beta Treatment_{cst} + \gamma \ln Y_{c0} + \epsilon_{cst} \tag{5.1}$$

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 the two year study period that circle i experienced all three staff as being paid under the scheme. $Y_{cs0}$is the value of the outcome variable at baseline (i.e. in the year prior to randomization). $Treatment$ is instrumented by a binary variable that represents the circle's randomization status. We include stratum fixed effects (α_s) given the lottery was stratified by these strata. As discussed above, all regressions 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, i.e. the maximum set of circles that have been involved together in a set of splits and merges since randomization.

To estimate the impact of the separate sub-treatments, we estimate:

$$\ln Y_{cst} = \alpha_s + \beta_1 Revenue_{cst} + \beta_2 RevenuePlus_{cst} + \beta_3 FlexibleBonus_{cst} + \gamma \ln Y_{c0} + \epsilon_{cst} \tag{5.2}$$

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 $\beta_3 = \frac{\beta_1+\beta_2}{2}$). We run analogous specifications when we examine the Information treatment.

For survey-based outcomes, we examine individual properties. As discussed above, we have two separate samples, the random sample of properties sampled from random GIS 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 random sample, we run regressions of the form:

$$Y_{ics} = \alpha_s + \beta Treatment_{ic} + \epsilon_{ic} \tag{5.3}$$

where i is an individual property. As above, we instrument for $Treatment$ with the randomization results. We include stratum fixed effects, and cluster standard errors at the circle level. In some

specifications, we also include controls for fixed property characteristics .[19]

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

$$Y_{ics} = \alpha_c + \beta_1 Treatment_{ic} * ReAssess_{ic} + \beta_2 ReAssess_{ic} + \epsilon_{ic} \tag{5.4}$$

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 $ReAssess$). Note that unlike equation (5.3) we now include circle-fixed effects (α_c), to capture fixed differences among circles between properties. We examine the analogue of equation (5.2) when we examine sub-treatments.

In interpreting equation (5.4), 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 - a extensive margin effect (i.e. the type/number of properties revalued can be impacted) and an intensive margin effect (a give reassessed property may now be dealt with differently) . For example, if Y_{ics} is the amount of bribes paid, the coefficient β_1 in equation (5.4) 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. the average bribe amount changes as the set/type of people who collude changes and conditional on collusion, the bribe amount changes). To shed some light on these effects , in Section 6.3 we will also examine how the composition of those in the re-assessed sample changes as well by estimating equation (5.4) on fixed characteristics of re-assessed properties.

6 Results

In Section 6.1, we first examine the overall impacts of the performance pay schemes on the key outcomes of interest, both revenue outcomes (current year and arrears revenue collected) and non-revenue incomes (accuracy of assessment and customer satisfaction with the tax department). We first examine the impact of being in any one of pay-for-performance schemes compared to controls, and then examine whether these schemes had differential impacts on these revenue and non-revenue outcomes.

[19] 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 **??** and **??**.

Section 6.2 then 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 the main results focus on the pay-for-performance aspect of the schemes, (i.e., price effects), Section 6.4 then considers a variety of alternative explanations for the results, such as perceptions of additional monitoring, income effects, and interactions with supervisors.

6.1 Main impacts

Impacts on Revenue Outcomes

Table 2 considers the impact of the performance pay schemes on (log) revenue at the end of each of the two years of the study. We consider 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 (i.e. collections of taxes from previous years that had yet to be paid). 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 for log arrears revenue are so much larger than for log current year revenue. 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. Focusing on Panel A, column (1) shows that compared to controls, revenue increased by 8.4 log points (i.e 8.7%) in treatment circles in the first year, and column (4) shows an increase of 8.6 log points in the second year. Examining the effects separately by current and arrears revenue, we find that the impact on current year revenue collection, is 6.7 log points (6.9%) in Year 1 and rises to 8.3 log points in Year 2. In contrast, there is a 15.6 percent increase in arrears revenue in Year 1 which falls to 10.7 percent (and is no longer statistically significant) in Year 2. Although these changes over the years are not statistically distinguishable the point estimates suggest that 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 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, Column (5) shows that by the end of Year 2, Revenue Circles collected 16.0 log points (17%) more revenue than control circles, compared to a 9.1 log point increase in Revenue Plus Circles and a 4.1 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 able to reject equality (p-value 0.075).

Impacts on Non-Revenue Outcomes

Despite showing substantial gains on revenue collection, performance pay schemes may still be undesirable if they reduced the accuracy of taxation or decreased the legitimacy of the tax department. In fact, as noted above, these are reasons why governments may have historically moved away from such schemes in tax collection. To test whether this is the case, we examine the impact of the treatments on measures of taxpayer satisfaction and accuracy of tax assessment, using property-level survey data. 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.

The first two measures are quality and satisfaction, in which we asked the respondent how they either rated the "quality of service" of the tax department or how "satisfied" they were with their service.[20] Answers were registered on a 5-point scale (though in the regressions these values have been rescaled to range from 0 to 1, with 1 as the highest value). Panel A of Table 3 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.

The zero results on quality and satisfaction are quite robust. In particular, we also show in the Online Appendix that the results are qualitatively unchanged if we control for observable property characteristics. Since perceptions often contain a substantial component that is a fixed attribute of the respondent, we also measured similar satisfaction questions for electricity bureau. Though controlling for perceptions of electricity reduces the standard error, it does not qualitatively affect the results. [21]

The lack of result overall, however, masks an important difference between the sub-treatments. In particular, examining Panel B for both quality and satisfaction, the point estimates for Revenue and Flexible Bonus are negative (-0.002 and -0.011 for Revenue; -0.064 and -0.056 for Flexible Bonus), whereas the point estimates for Revenue Plus are positive (0.032 and 0.023). While only the Flexible Bonus coefficients are statistically distinguishable from 0 at conventional levels when compared to controls, the schemes are distinguishable from each other. That is, we can reject the null of equal effects for the three schemes with a p-value of 0.011 for quality and 0.057 for satisfaction.[22] They are also both jointly significant.

We next turn to the second main non-revenue dimension, the inaccuracy of assessment of the property. We focus on the Gross Annual Rental Value ($GARV$) of the property, which is the main

[20] See footnote 7 for exact question wording.

[21] Online Appendix Tables also shows that the results are qualitatively similar if we use ordered probit models instead of the linearized variable with OLS.

[22] Note that this is also true when we include a full vector of controls (Online Appendix table).

measure of a property's tax value, before exemptions and reductions are applied.[23] The inaccuracy of assessment is calculated by comparing the inspector's assessment with the assessment we compute using our independent survey metric, normalized by the sum of the two measures, according to the following formula:

$$Inaccuracy = \frac{|GARV_{Inspector} - GARV_{Survey}|}{(GARV_{Inspector} + GARV_{Survey})} \tag{6.1}$$

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

$$TaxGap = \frac{GARV_{Inspector} - GARV_{Survey}}{(GARV_{Inspector} + GARV_{Survey})} \tag{6.2}$$

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

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 increase 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) 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 downsides in terms of taxpayer satisfaction and the accuracy of tax perceptions for the typical property. Second, multi-dimensional incentives appear to matter. Comparing the Revenue and Revenue Plus scheme, we find that the Revenue scheme increased current revenue by about 14 percent whereas the Revenue Plus scheme increased current revenue by only about 10.5 percent; on the other hand, customer satisfaction was slightly higher in the Revenue Plus scheme. This suggests that although the effects on tax payer satisfaction in Revenue were small and not statistically distinguishable from controls, to the extent one was worried about these issues a multi-dimensional incentive scheme could potentially address them, albeit with substantial revenue costs. The Flexible Bonus scheme did worse on all dimensions measured here than either Revenue or

[23] 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 continuou function of the underlying property characteristics and hence is much more robust to measurement error.

Revenue Plus. This provides suggestive evidence against subjective and more multi-dimensional assessments and in favor of clearer, formulaic based assessments that consider fewer metrics. This may be especially so in contexts where there may be concern about how the more complex and flexible assessments may be applied, as suggested in the theoretical work on subjective bonuses (Baker, Gibbons and Murphy, 1994; Prendergast and Topel, 1996; Prendergast, 1999; Bentley, 2003).

6.2 Decomposing the Revenue Impact

To understand how the revenue changes occurred, we can decompose them further using the administrative data. There are three margins that could be affected: the tax-base before exemptions are granted, 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:

$$Revenue = TaxBase * NonExemptionRate * RecoveryRate \tag{6.3}$$

where $NonExemptionRate = \frac{TaxBaseAfterExemptions}{TaxBaseBeforeExemptions}$ and $RecoveryRate = \frac{Revenue}{TaxBaseAfterExemptions}$. Taking logs of equation 6.3we obtain an expression that additively decomposes the source of the tax revenue.

Table 4 reports the results of estimating equation (5.1) for this decomposition exercise for both treatment years. We control for the baseline values of $\ln TaxBase$, $\ln NonExemptionRate$, and $\ln TaxBase$ in all columns. Columns (1) and (5) begin by reproducing the same regressions of $\ln Revenue$ from Columns (1) and (4) of Table 2, with the addition of baseline $\ln NonExemptionRate$ and $\ln TaxBase$ as controls. (The sample size is reduced slightly given a few missing observations for $\ln NonExemptionRate$ and $\ln TaxBase$ at baseline.) 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.[24]

Appendix Table A.4 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 were had multi-dimensional incentives, inspectors focused more on the tax base, at least as we can measure

[24]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 , by year 2 there is therefore likely not much room left to improve the arrears tax base.

it here.

6.3 Understanding Channels: Tax Base Changes

The results above suggest that property tax revaluations, the mechanisms by which the tax base is increased, plays a critical role in driving the treatment effect. These revaluations can take two forms: previously untaxed properties added to the tax rolls for the first time, or previously taxed properties whose valuations are being updated.

The conceptual framework in Section 3 illustrates how taxpayer and tax collectors may collude not to pay taxes, a likely scenario in the context we study. For the purposes our our treatments, the framework shows how performance pay can make such collusions 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 in much more detail how the treatments shift the reassessment margin. For this data, we use data from the underlying tax registers, reflecting which properties have in fact been re-assessed.

How many are re-assessed?

We begin by examining how the re-assessment margin changes the number and composition of properties that are re-assessed. Panel A of Table 5 shows the total number of re-assessed properties, broken down by properties assessed for the first time and those who had previously been on the tax rolls but whose valuation was updated. For these data, our staff counted the number of properties added to the tax rolls or reassessed from the underlying tax registers, so this reflects the actual tax base as recorded by the government, and is not subject to manipulation by tax staff in totaling. We control for the value of the dependent variable from the baseline year (i.e. 2010-2011) to capture heterogeneity across circles in their underlying rate of change of properties.

The results show a substantial increase in the number of re-assessments. On average, (over the two year treatment period) there are 82 more properties with new or updated valuations in treatment areas compared to controls, about a 85% increase over the control group. Most of this increase comes from newly added properties. Column 2 shows that treatment circles add about 73 more newly valued properties to the tax rolls than controls (197% 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: 73 new properties represents about 3 percent of the average number of properties in the circle.

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 (5.4). 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 on these characteristics in control circles; the interaction term $Reassess \times Treatment$ captures how the 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 being taxed 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 5, where Panel B examines characteristics of the property and Panel C examines characteristics of the owner. In general, in both treatment and control areas, Panel B shows that re-assessed properties look to be those that are subject to higher tax rates than typical property. For example, re-assessed properties are substantially more valuable: 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 69% higher than the mean property in control areas. They also have more floors, are more likely to have been recently renovated, to belong to a more expensive tax bracket (tax category), and to be commercial (which is taxed at a higher rate) and 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 56% larger the average re-assessed property in control areas (i.e. therefore reassessed properties in treatment areas have a 125% higher GARV than the typical property in control areas from the random sample). They also focus more on commercial rented properties, which have the highest assessments per square foot of area.

Panel C considers changes in owner characteristics. The one striking finding is that those owners who property 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 reversed in treatment areas – so that while connected owners seem to enjoy an advantage in general, this is undone in treatment areas. Thus the performance pay incentives induce tax collectors to "move down" along influence curve - i.e. at baseline the less influential are more likely to be reassessed but then as one needs to reassess more, there is no choice for the inspector but to start reassessing the somewhat more influential. 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 with potentially influential owners.

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.[25] For properties that were in the non-collusive equilibrium before, and remain there, we would expect no changes.

To investigate these heterogeneous effects, we again estimate equation (5.4), which examines the difference between what happens to regular 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 (5.4), i.e. the coefficient on $ReAssess$, captures how, in both treatment and control, those properties who are reassessed differ from the random sample of properties. Since equation (5.4) includes circle fixed effects, we also estimate equation (5.3) on the random sample of properties to capture how random properties in treatment areas differ on these variables compared to equivalent properties in control areas.

Panel A of begins with the estimation of the treatment effect on the random sample of properties, from equation (5.3). Since only about 8 percent of randomly selected properties have been re-assessed over the treatment period, we do not have sufficient sample to detect the tax increase we find in treatment areas if that stems from these reassessed properties. We find that, for the random sample of properties, tax payments are essentially unchanged. Bribe payments, 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, increase substantially, by 583 Rupees (US \$6, about 30 percent higher than for the average control area property).[26] The frequency of bribe payments also increases substantially. The one metric of corruption that does not change is the overall perceptions of corruption in the tax department.

Panel B of Table 6 estimates how these effects differ for re-assessed properties from estimating equation (5.4). There are several key results to note. First, compared with non-reassessed properties, re-assessed properties in control circles pay substantially higher taxes – about Rs 3400, or about 73 higher than the control group mean for random properties in Panel A.[27] This is even

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

[26]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.

[27]Note that to be consistent with the main results in Table 2, for tax payment we include the baseline value as a control. No other columns have control variables.

more true in treatment areas, where re-assessed properties pay an additional 2,200 more than non-reassessed properties.[28] On the other hand, the increase seen in bribes treatment areas is not seen for re-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.[29]

Table 7 examines how how re-assesed properties differ from other properties in terms of the key non-revenue metrics we examined in Table 3 – satisfaction, inaccuracy and the tax gap. It is interesting to note that re-assessed properties are not, broadly speaking, unsatisfied with the tax department. In fact, Table 7 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 reassessed and switch between the collusive and non-collusive equilibrium in response to the treatment are those who are closest to being indifferent between the two regimes. The switch from collusive to non-collusive equilibrium may therefore represent a second-order utility change for these property owners, even though it yields a first-order change in revenue for the government.

The results are striking when we look at 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 in general, randomly-selected properties in the control group, are on average 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 0.186); i.e. re-assessed properties are on average taxed on average at the amount our independent survey team would predict.) These effects are similar in both treatment and control areas. These results confirm

[28] The results here are consistent with the 8% 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.

$$E[TaxPayment] = E[TaxPayment|Reassessed]P(Reassessed)+E[TaxPayment|Non-reassessed]P(Non-reassessed)$$

Based on our estimates here and data on reassessment rates (8% of taxable properties were reassessed in the cumulative two year treatment period in control circles and for simplicity we treat our random sample as composed only of non-reassessed properties), this average in control areas is

$$(0.08)(4245 + 3419) + (0.92)(4245) = 4518.52$$

This gives an average tax per property of Rs 4245 in control areas. Using our treatment effect estimates (i.e. increases in the number of reassessed properties and the great payments received from such properties and the effectively unchanged payments for non-reassessed properties), the analogous average tax in treatment circles is given by

$$(0.118)(4245 + 3419 + 2204) + (0.882)(4245) = 4908.514$$

An increase in the average tax per property from Rs. 4,518 to Rs. 4,908 represents a 8.6% increase in tax collection which is quite close to the observed effect from our admin data of around 9% (8.6 log points)

[29] Appendix Table A.7 repeats analysis of Table 6 broken down by the three subtreatments. The results do now show substantial differences in these dimensions among the three subtreatments.

the view of re-assessment as a bargaining breakdown: unlike typical randomly-selected properties, which in general are undertaxed, re-assessed properties are assessed more accurately and are neither over- nor under-taxed on average.

Combined, the results here paint a picture consistent with the theoretical framework: most properties pay no more taxes but 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 Alternative Explanations

We have thus far interpreted our results as a result of changes in collusion due to the increased marginal incentives provided to collect more taxes. However, the schemes we introduced also have other aspects that could also enhance performance. First, they receive slightly more salient information and may be potentially 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, while collusion is clearly part of the story given the bribe results, it may also be accompanied by more effort by the inspectors in the field to gather more up-to-date information about property status and hence it's tax liability. Finally, supervisors might change their behavior in response to the incentives. This section explores each of these additional affects in turn, and concludes that the price effects of the incentives seem to be the primary way in which the incentives had an impact.

Information and monitoring effects

In addition to receiving the performance incentives, treatment inspectors also receive additional information. Specifically, each quarter they receive a form explaining how their payments were calculated, which shows their revenue collected and the corresponding benchmarks. While this is not new information per se, since all of this information is calculated based on historical information they themselves previously reported, it is presented in a slightly different way. Moreover, the fact that staff were brought together every quarter for these meetings (and received trainings initially and then were quizzed about their understanding of the schemes etc. subsequently) could have made their performance more salient, or given them the impression that they were being more intensively monitored.

We test for these effects in two ways. First, to test for the impact of information and perceptions of monitoring, starting in year 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 purely isolates the impact of information & perceived monitoring per se. We confirmed

through a follow-up survey with inspectors that they understood the scheme and knew that they would not be paid that year though they could potentially have held beliefs about their chances of being selected in subsequent years (see Appendix Table A.3).

While in our baseline specifications we included these circles as part of the control group , in Table 8 we separate out the information scheme and report tests of equality compared to the other schemes. The results in Table 8 show generally positive point estimates associated with the information scheme (particularly in the arrears treatment) but they are not statistically distinguishable from zero. For total revenue, the point estimate is that the information scheme is associated with 5.8 percentage points higher revenue, compared with 13.5 percentage points for the Revenue scheme (a test for equality has a p-value of 0.186); for current-year revenue, the point estimate is that the information scheme is associated with 4.1 percentage points higher revenue, compared with 16.0 percentage points for the Revenue scheme (a test for equality has a p-value of 0.101). Taken literally, these results that something between one-quarter and one-third of the total effect of the Revenue scheme could be attributable to the information component (although we cannot reject that the information-only effect was zero), but that the vast majority of the effect was attributable to the financial aspect of the performance incentives.

Second, to examine whether inspectors in treatment schemes believed they were being monitored, we asked them in a survey at the end of the second year about their perceived levels of pressure from supervisors and levels of monitoring. The results are shown in columns (1) and (2) of Table 9. Panel A of Appendix Table 9 shows that there is, on average, no difference in perceived pressure or monitoring between treatment and control areas. Panel B shows that, if anything, the Revenue scheme inspectors report less monitoring, not more. The only place we see more is that the Flexible Bonus inspectors report more pressure from supervisors, which is consistent with the fact that they are being subjectively evaluated by their supervisors. On net, the lack of any increase in perceived monitoring, and the fact that the Revenue treatment – which shows the largest effects – if anything has less monitoring – further confirms that perceived monitoring by supervisors is unlikely to be the main effect driving the effects in the paper.

Income effects

In addition to a price effect (e.g., the derivative of an inspector's earnings with respect to tax collections increases with performance pay), performance pay may also have income effects (e.g., staff in the performance pay treatment earn higher income because the bonus is paid in addition to normal wages). We have already discussed above how price effects can have an impact (as in the model in Section 3 or a more traditional moral-hazard model such as Hölmstrom 1979), but it important to consider the possibility of income effects as well. For example, one could imagine that honesty is a normal good (i.e. inspectors take bribes because they have a high marginal utility of income).

There are several pieces of evidence that suggest that the impacts observed are not being primarily driven by income effects. First, all three of the performance-pay schemes were designed to generate approximately similar expected income (and indeed did so), yet we saw above that the generate very different impacts. For example, the reward rate formula based on tax collected is identical for the Revenue and Revenue plus schemes, whereas the "Plus" part is, by construction, designed to have zero expected additional earning (staff are grouped into terciles where the middle group gets no additional income and the top group receives exactly the same additional bonus that the bottom group loses). In practice, the mean earnings in all three schemes were quite comparable in both years, and if anything, favored schemes that saw smaller increases in revenue 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. Despite the fact that mean earnings were similar, the schemes produced substantially different outcomes: the increase in tax revenue in Revenue was almost double what it 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.

A second approach is to use the lag structure of the incentive scheme. Specifically, recall from equation (4.1) that in the Revenue and Revenue Plus scheme, incentives were paid on revenues above benchmarks. Benchmarks, in turn, we determined based on the 2nd, 3rd, and 4th lags of revenue, but not the 1st lag.[30] However, revenue can be closely approximated by an AR(1) process. Specifically, if one regresses 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.[31]

This feature of the design suggests a way of identifying income effects. Specifically, we regress

$$LnRevenue_t = \gamma_1 LnRevenue_{t-1} + \gamma_2 LnBenchmark_{t-2} + \epsilon \tag{6.4}$$

and form the prediction $\hat{LnRevenue_t}$. This is the amount of revenue that would be collected under business-as-usual. An inspector in the Revenue or Revenue Plus group would expect to earn

$$IncomeShock_t = \alpha \left(\hat{Revenue_t} - Benchmark_{t-2} \right) \tag{6.5}$$

[30]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). Logitistically, 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.

[31]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.

simply from business-as-usual. That is, equation (6.5) represents the pure income shock that an inspector randomized into Revenue or Revenue Plus would receive compared to an inspector in the control group. As long as all inspectors are above the benchmark, and so face the same price effects on the margin, this represents a pure income shock. 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.[32] Since $IncomeShock_t$ is defined in both treatment 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:

$$\begin{aligned} LnRevenue_{ct} &= \beta_1 Treatment_c + \beta_2 LnIncomeShock_{ct} \times Treatment_c + \beta_c IncomeShock_{ct} \\ &\quad + \beta_4 LnRevenue_{ct-1} + \beta_5 LnBenchmark_{ct-2} + \epsilon \end{aligned} \tag{6.6}$$

The key coefficient of interest is β_2.[33] Note also that this only works for the first year of the program (2012), in the second year the first lag of revenue is the revenue realized in the first treatment year, which is endogenous. We estimate this both separately by treatment and then pooling Revenue and Revenue plus together.

Table 10 presents the results. As is shown in Table 10, there is no difference in performance based on the infra-marginal component of the revenue treatments. Again, this suggests that the key component is the price effect, not the income effect..[34]

Changes in 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 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, in this section we try and directly examine whether there

[32] 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 (6.4) and (6.5) for current and arrears and then add to get the total income shock.

[33] 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.

[34] An approach to directly examine price effects is to use 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).

is some direct evidence for the effort margin. We should note though that this effort could also be consistent with a pure collusion model since the tax collector may require greater rebargaining effort in such cases as well.

At the end of the two year experiment (June 2013), we interviewed inspectors and asked them about both their work effort in terms of time use and time allocation (this is the same survey in which we asked about supervisory pressure). The results are shown in Table 9. The first order-implication is that there is little observable change in effort – there is no change in total hours spent working per day. 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.

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, we introduced such supervisory performance-pay. 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 Assisstant ETOs. We randomized at the level of ETO treated all AETOs who worked underneath them. Supervisors are responsible for not only 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. 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[35].

Table 11 reports two sets of results. In column 1, we examine the overall average effect of the supervisor incentives. 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. In column 2, we investigate 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 affects orthogonal to one other. The results suggest that there may be some positive interactions between circle and supervisory staff both being paid for performance and that if anything, paying supervisors only may in fact be detrimental to overall collections. We should caution that while we may be

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

tempted to do so, given the imprecise estimates on the supervisory treatments due to lower sample size, it may be more prudent not to make too much of this effect.

7 Conclusion

Our paper shows that performance pay mechanisms can be quite effective in raising additional taxes while generating little political and performance costs that were 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 might such costs 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 present compelling evidence that there is - 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? The historical tax farming literature suggests that tax collectors may have been more oblivious (than the central authority) to the potential political costs they imposed when raising taxes. However, tax collectors in our context may not be as free to do so - they are not locally powerful enough to be relatively unaffected by the displeasure of the population they tax. In fact, more often then not they may have weaker socioeconomic and political influence. To the extent that this is the case, they may also be concerned about the potential political costs 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 they individual tax-payer may be quite influential and/or because they may collectively be powerful (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 given the general level of dissatisfaction tax-payers would have about how their taxes are utilized (locally) by the state.

So how might a tax-collector balance their increased incentives to raise more taxes due to performance-pay schemes with potentially not generating resulting taxpayer dissatisfaction? One might have imagined two different potential responses. One response might have been 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 have instead focused their efforts on a small number of high value owners. This would return the largest return per property, and avoided alienating a large number of people, but could have been risky if they alienated 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. While we do not know whether this was the best option to balance their goal of raising more taxes without inducing excessive dissatisfaction, our results do suggest that it was fairly effective.

References

Bahl, Roy, Sally Wallace and Musharraf Cyan. 2008. Pakistan: Provincial government taxation. Technical report International Center for Public Policy, Andrew Young School of Policy Studies, Georgia State University.

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

Bartlett, Bruce. 1994. "How excessive government killed ancient Rome." *Cato J.* 14:287.

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

Carillo, Paul, Dina Pomeranz and Monica Singhal. 2014. Tax Me if You Can: Evidence on Firm Misreporting Behavior and Evasion Substitution. Technical report Harvard Kennedy School.

Fisman, Raymond and Shang-Jin Wei. 2004. "Tax Rates and Tax Evasion: Evidence from 'Missing Imports' in China." *Journal of Political Economy* 112(2):471–500.

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

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

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

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

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115(2):200–249.

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

Piracha, Mujtaba and Mick Moore. 2013. "The Informality of Formal Tax Collection in Pakistan.".

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

Prendergast, Canice and Robert H Topel. 1996. "Favoritism in Organizations." *The Journal of Political Economy* 104(5):958–978.

White, Eugene N. 2004. "From privatized to government-administered tax collection: tax farming in eighteenth-century France1." *The Economic History Review* 57(4):636–663.

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	68
Information	0	70	0	66
Control	322	194	338	212

Circles that did not consent were assigned treatment values of 1/3 for each treatment type. Values are rounded.

Table 2: 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	0.084***	0.067**	0.145**	0.086***	0.083***	0.102
	(0.027)	(0.027)	(0.068)	(0.030)	(0.031)	(0.082)
Panel B: Subtreatments						
Revenue	0.110***	0.101***	0.125	0.122***	0.145***	-0.004
	(0.035)	(0.035)	(0.099)	(0.043)	(0.044)	(0.133)
Revenue Plus	0.077	0.083*	0.068	0.087**	0.076*	0.165
	(0.048)	(0.050)	(0.108)	(0.042)	(0.044)	(0.111)
Flexible Bonus	0.062*	0.015	0.234**	0.047	0.027	0.134
	(0.037)	(0.035)	(0.098)	(0.040)	(0.041)	(0.104)
N	481	481	481	482	482	479
Mean of control group	15.672	15.380	14.031	15.746	15.519	13.916
Rev. vs. Multitasking p.	0.332	0.209	0.817	0.224	0.045	0.266
Objective vs. Subjective p.	0.470	0.072	0.221	0.182	0.064	0.650
Equality of Schemes	0.568	0.136	0.441	0.328	0.077	0.530
Joint significance	0.009	0.018	0.092	0.019	0.008	0.342

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table 3: 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	0.005
	(0.022)	(0.022)	(0.012)	(0.021)
Panel B: Subtreatments				
Revenue	0.001	-0.010	0.009	-0.031
	(0.036)	(0.038)	(0.017)	(0.029)
Revenue Plus	0.035	0.025	0.031**	0.011
	(0.025)	(0.026)	(0.016)	(0.031)
Flexible Bonus	-0.061**	-0.054*	-0.012	0.024
	(0.029)	(0.030)	(0.017)	(0.030)
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.	0.710	0.905	0.955	0.123
Objective vs. Subjective p.	0.014	0.062	0.078	0.280
Equality of Schemes	0.011	0.056	0.096	0.278
Joint significance	0.028	0.118	0.134	0.464

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table 4: 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	0.069***	0.082***	-0.032	0.019	0.082***	0.045	0.007	0.029
	(0.026)	(0.029)	(0.022)	(0.025)	(0.029)	(0.033)	(0.022)	(0.021)
Current	0.066**	0.077***	0.001	-0.012	0.088***	0.060**	0.018	0.011
	(0.027)	(0.027)	(0.014)	(0.022)	(0.031)	(0.029)	(0.016)	(0.016)
Arrears	0.103	0.124*	-0.057	0.038	0.065	-0.016	0.042	0.039
	(0.065)	(0.070)	(0.046)	(0.043)	(0.079)	(0.088)	(0.051)	(0.045)
N (total)	473	470	470	473	474	474	474	474
Mean of control group	15.681	16.115	-0.188	-0.239	15.757	16.151	-0.218	-0.176

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table 5: Selection Effects in Reassessment

Panel A

	Total Number of Section 9 Properties Added to Tax Rolls in Treatment Period	Number of New Properties Added to Tax Rolls in Treatment Period	Number of Reassessed Properties Added to Tax Rolls in Treatment Period
Treatment	81.647*	72.756**	8.891
	(44.267)	(33.550)	(21.971)
N	234	234	234
Mean of control group	97	37	60

Panel B

				Components of GARV						
	GARV	Number of floors	Last renovation was ≤ 2 years ago	Land area (sq. feet)	Total covered area, all uses	Main Road	Category	Percent of property commercial	Percent of property commercial and rented	Tax Liability
Re-assess * Treatment	20302.579	0.002	-0.005	-266.635	854.049	-0.002	-0.216***	0.018	0.074**	4044.323
	(16190.384)	(0.049)	(0.020)	(732.635)	(755.810)	(0.047)	(0.082)	(0.036)	(0.029)	(3538.198)
Re-assess dummy	24773.582***	0.078***	0.095***	336.275	-206.893	0.064***	0.205***	0.216***	0.176***	5496.217***
	(7840.168)	(0.026)	(0.011)	(518.375)	(378.620)	(0.024)	(0.042)	(0.019)	(0.015)	(1730.354)
N	15489	16352	16128	16352	16346	16352	15489	16226	16227	15489
Mean of control group in random sample	35986.470	1.570	0.018	2703.994	2803.921	0.456	3.759	0.345	0.166	6483.796

Panel C

	Approximate age of owner	Per-capita wages	Predicted expenditure given assets	Connected to Politician	Connected to Politician/ Government/ Police
Re-assess * Treatment	-0.342	-806.466	109.030	0.021*	0.005
	(0.782)	(1057.937)	(208.608)	(0.012)	(0.027)
Re-assess dummy	-0.655	17.263	-95.116	-0.013**	0.005
	(0.400)	(513.033)	(123.203)	(0.006)	(0.014)
N	13406	13765	13954	16354	16354
Mean of control group in random sample	50.702	16281.547	6291.640	0.053	0.356

Regressions for Panel A control for number of new and reassessed properties added in FY 2011.
Regressions for Panels B and C include circle-level fixed effects
Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

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

	(1) Self-reported Tax Payment	(2) Bribe Payment	(3) Frequency of Bribe Payment	(4) Perception of Corruption	(5) Total payment to tax dept. (Tax + Bribe)
Panel A: Random Sample					
Treatment	-120.1	583.6*	.1985**	.0111	-324.9
	(309.4)	(327.3)	(.0929)	(.025)	(492.5)
N	9632	5993	4802	6050	5593
Mean of control group	4919.067	1874.542	0.683	0.644	6790.628
Panel B: Re-assessed					
Re-assessed * Treatment	2204*	-546.4	-.1559*	-.0031	2560
	(1288)	(360.8)	(.092)	(.0219)	(1790)
Re-assessed	3419***	-64.81	.0142	-.0191*	3635***
	(693.8)	(178)	(.0405)	(.0108)	(991.7)
N	13693	8207	6993	8268	8203
Sample	Full	Phase 1	Phase 1	Phase 1	Phase 1
Mean of control group in random sample	4713.484	1874.542	0.683	0.644	6446.671

Robust standard errors, clustered by circle, in parentheses
Specification in Columns 1 and 5 of Panel A controls for self-reported tax at baseline.
* p<0.10, ** p<0.05, *** p<0.01

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

	(1) Quality	(2) Satisfaction	(3) Inaccuracy	(4) Tax Gap
Re-assessed * Treatment	0.008	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 random sample	0.538	0.555	0.339	-0.103

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table 8: Information effects

	Total	Current	Arrears
Revenue	0.135***	0.160***	0.044
	(0.045)	(0.046)	(0.138)
Revenue Plus	0.100**	0.091**	0.213*
	(0.044)	(0.046)	(0.117)
Flexible Bonus	0.060	0.041	0.181
	(0.041)	(0.043)	(0.110)
Information	0.058	0.065	0.217
	(0.052)	(0.050)	(0.137)
N	482	482	479
Mean of dep. var.	15.753	15.524	13.909
Rev. = Info, p.	0.186	0.101	0.293
Rev+ = Info, p.	0.464	0.655	0.978
Flex. Bonus = Info, p.	0.973	0.671	0.805
All Treatments = Info, p.	0.411	0.503	0.577

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table 9: Inspector Monitoring and Effort

	(1) Pressure from supervisors	(2) Level of monitoring	(3) Team Effort	(4) Hours/Day spent in field	(5) Hours/Day spent in office	(6) Total hrs worked in typical day
Panel A: Main Treatment						
Treatment	.0177	-.0253	-.566	-.278**	.176	-.102
	(.0297)	(.0428)	(1.99)	(.138)	(.143)	(.0728)
Panel B: Subtreatments						
Revenue	.0218	-.128*	-1.89	-.0776	.0161	-.0615
	(.048)	(.0717)	(3.78)	(.213)	(.214)	(.12)
Revenue Plus	-.0525*	.0233	3.05	-.284	.0924	-.191***
	(.0289)	(.0601)	(2.66)	(.178)	(.186)	(.0699)
Flexible Bonus	.0886*	-.0272	-1.1	-.423**	.297	-.126
	(.0489)	(.0557)	(2.68)	(.202)	(.192)	(.0875)
N	352	351	352	352	352	352
Mean of control group	.104	.775	90.6	5.46	2.68	8.14

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table 10: Income Effects

	(1)	(2)	(3)	(4)
Panel A: Ln(Total Revenue)				
Objective * Income Shock	0.000	0.008		
	(0.000)	(0.017)		
Revenue * Income Shock			0.000	0.003
			(0.000)	(0.029)
Revenue Plus * Income Shock			0.000	0.007
			(0.000)	(0.022)
Flexible Bonus * Income Shock	-0.000	-0.027	0.000	-0.027
	(0.000)	(0.021)	(0.000)	(0.021)
Income Shock	-0.023	-0.016**	-0.021	-0.016**
	(0.021)	(0.007)	(0.021)	(0.007)
Panel B: Ln(Current Revenue)				
Objective * Income Shock	0.000	0.017		
	(0.000)	(0.017)		
Revenue * Income Shock			0.000	0.008
			(0.000)	(0.028)
Revenue Plus * Income Shock			0.000	0.019
			(0.000)	(0.022)
Flexible Bonus * Income Shock	0.000	0.004	0.000	0.004
	(0.000)	(0.016)	(0.000)	(0.016)
Income Shock	-0.004	-0.004	-0.003	-0.005
	(0.021)	(0.006)	(0.021)	(0.006)
Panel C: Ln(Arrears Revenue)				
Objective * Income Shock	0.000	-0.032		
	(0.000)	(0.052)		
Revenue * Income Shock			0.000	0.081
			(0.000)	(0.068)
Revenue Plus * Income Shock			-0.000	-0.091*
			(0.000)	(0.054)
Flexible Bonus * Income Shock	-0.000	-0.069	-0.000	-0.068
	(0.000)	(0.070)	(0.000)	(0.070)
Income Shock	0.091*	0.008	0.091*	0.010
	(0.054)	(0.018)	(0.054)	(0.018)
N	481	481	481	481
Income Shock in	Level (in Million Rs.)	Log	Level (in Million Rs.)	Log

Robust standard errors, clustered by circle, in parentheses.
Objective treatments are Revenue and Revenue Plus.
* p<0.10, ** p<0.05, *** p<0.01

Table 11: Supervisory Treatment

	(1)	(2)
Panel A: Total Revenue		
Supervisory treatment	-0.067	-0.086*
	[0.262]	[0.095]
Circle treatment		0.058
		[0.170]
Circle Treatment * Supervisory Treatment		0.094
		[0.255]
Panel B: Current Revenue		
Supervisory treatment	-0.108	-0.129**
	[0.163]	[0.036]
Circle treatment		0.063
		[0.161]
Circle Treatment * Supervisory Treatment		0.068
		[0.409]
Panel C: Arrears Revenue		
Supervisory treatment	-0.035	-0.106
	[0.807]	[0.470]
Circle treatment		0.029
		[0.788]
Circle Treatment * Supervisory Treatment		0.252
		[0.273]
N	479	479

Randomization inference p-values in brackets.
* p<0.10, ** p<0.05, *** p<0.01

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	-0.0342	0.0279	-0.0521	-0.0509	15.46	0.0501	15.65	-0.167
		(0.0416)	(0.0593)	(0.0569)	(0.0584)		(0.0648)		(0.0887)
Log Recovery Rate	-0.333	-0.0148	0.00713	-0.00192	-0.0388	-0.329	-0.0263	-0.366	0.00645
		(0.0245)	(0.0348)	(0.0335)	(0.0404)		(0.0374)		(0.0407)
Log Non-exemption Rate	-0.251	-0.0245	-0.00158	0.00881	-0.0590	-0.245	-0.0352	-0.267	0.00898
		(0.0191)	(0.0210)	(0.0232)	(0.0382)		(0.0350)		(0.0193)
Number of staff posted	2.564	0.0546	0.0378	0.0556	0.0877	2.576	-0.0772	2.549	0.0544
		(0.0527)	(0.0695)	(0.0756)	(0.0775)		(0.0802)		(0.0621)
All positions filled	0.519	0.0591	0.0431	0.0943	0.0545	0.531	-0.0724	0.538	-0.0113
		(0.0443)	(0.0651)	(0.0636)	(0.0658)		(0.0646)		(0.0562)
Log benchmark	15.44	-0.0175	0.0357	0.0141	-0.0730	15.44	-0.00954	15.59	-0.114
		(0.0442)	(0.0623)	(0.0587)	(0.0644)		(0.0738)		(0.0951)
FY 10-11 log growth rate	0.0277	0.0129	0.0250	0.00960	0.00297	0.0216	0.0340	0.0259	-0.00685
		(0.0139)	(0.0162)	(0.0182)	(0.0240)		(0.0227)		(0.0195)
P-val, joint sig.		0.280	0.670	0.00174	0.492		0.268		0.00295
P-val, from RI		0.236	0.676	0.00400	0.601		0.420		0.132

Table A.2: Summary Statistics

	Mean	SD	N
Panel A: Administrative Data			
Log Tax Base (Current) FY 2012	15.770	0.709	477
Log Tax Base (Arrears) FY 2012	14.563	1.428	477
Log Tax Base (Total) FY 2012	16.125	0.824	477
Log Revenue (Current) FY 2012	15.372	0.718	481
Log Revenue (Arrears) FY 2012	14.071	1.211	481
Log Revenue (Total) FY 2012	15.675	0.751	481
Log Recovery Rate (Current) FY 2012	-0.203	0.189	481
Log Recovery Rate (Arrears) FY 2012	-0.293	0.372	481
Log Recovery Rate (Total) FY 2012	-0.241	0.245	481
Log Non-Exemption Rate (Current) FY 2012	-0.190	0.121	477
Log Non-Exemption Rate (Arrears) FY 2012	-0.192	0.361	477
Log Non-Exemption Rate (Total) FY 2012	-0.203	0.203	477
Log Tax Base (Current) FY 2013	15.861	0.733	482
Log Tax Base (Arrears) FY 2013	14.400	1.372	479
Log Tax Base (Total) FY 2013	16.143	0.806	482
Log Revenue (Current) FY 2013	15.524	0.728	482
Log Revenue (Arrears) FY 2013	13.909	1.170	479
Log Revenue (Total) FY 2013	15.753	0.738	482
Log Recovery Rate (Current) FY 2013	-0.143	0.141	482
Log Recovery Rate (Arrears) FY 2013	-0.226	0.335	479
Log Recovery Rate (Total) FY 2013	-0.169	0.193	482
Log Non-Exemption Rate (Current) FY 2013	-0.194	0.133	482
Log Non-Exemption Rate (Arrears) FY 2013	-0.265	0.435	479
Log Non-Exemption Rate (Total) FY 2013	-0.221	0.209	482
Panel B: Survey Data			
GARV	31914.97	248025.86	11186
Self-reported tax payment in FY 2013	4245.85	20254.83	10047
Self-reported tax payment in FY 2011	3011.17	18008.79	11584
Inaccuracy	0.34	0.27	9879
Tax Gap	-0.10	0.42	9879
Quality of Tax Department (0-1)	0.53	0.22	6050
Satisfaction with Tax Department (0-1)	0.55	0.23	6050
Degree of Corruption in Tax Department (0-1)	0.64	0.22	6050
Bribe Payment	2073.34	3931.64	5993
Frequency of Bribe Payment	0.76	0.88	4802
Tax + Bribe Payment	6167.80	25086.34	5989
Number of floors	1.58	0.66	12000
Last Renovation was $\leq$ 2 years ago	0.02	0.13	11820
Land Area (sq. ft.)	2509.62	6668.62	12000
Total Covered Area (sq. ft.)	2585.07	18536.38	12000
On Main Road	0.46	0.50	12000
Taxation category (1-7)	3.74	1.55	11186
Fraction of covered area - Commercial	0.35	0.42	11912
Fraction of covered area - Commercial and Rented	0.16	0.33	11913
Age of owner	50.68	11.37	9222
Per-capita wages	16565.74	16661.69	9459
Predicted values of expenditure given assets owned	6214.37	2963.93	9600
Connected to Politician	0.05	0.22	12000
Connected to Politician/Government/Police	0.35	0.48	12000
Panel C: Inspector Survey Data			
How would you score your circle team effort in your duties?	90.15	14.18	370
In a typical work day how many hours did you spend in the field	5.36	0.92	370
In a typical work day how many hours did you spend in the office	2.75	0.96	370
Total hours worked in a typical day (field + office)	8.10	0.50	370
Pressure from supervisors	0.10	0.20	370
Level of monitoring	0.77	0.29	369

Table A.3: Inspectors' Knowledge of Treatments

	Dependent variable: Treatment inspectors believe themselves to be in:			
	Revenue	Revenue Plus	Flexible Bonus	Information (placebo)
Independent variable: true treatment				
Revenue	1.262***	0.189**	0.024	-0.034
	(0.105)	(0.074)	(0.036)	(0.026)
Revenue Plus	0.067	1.262***	-0.002	-0.041*
	(0.054)	(0.085)	(0.011)	(0.022)
Flexible Bonus	-0.012	0.058	1.129***	-0.030
	(0.033)	(0.044)	(0.082)	(0.021)
Information Placebo	-0.040	-0.043*	0.002	1.363***
	(0.027)	(0.026)	(0.015)	(0.118)
N	352	352	352	352
Mean of control group	0.021	0.011	0.000	0.225
Rev. vs. Multitasking p.	0.000	0.000	0.000	0.953
Objective vs. Subjective p.	0.000	0.000	0.000	0.639
Equality of Schemes	0.000	0.000	0.000	0.826
Joint significance	0.000	0.000	0.000	0.321

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table A.4: Impacts on Tax Base and Recovery Rates, Sub 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
Panel A: Total								
Revenue	0.113***	0.080**	-0.021	0.053*	0.130***	0.040	0.035	0.055**
	(0.034)	(0.040)	(0.031)	(0.029)	(0.045)	(0.051)	(0.032)	(0.026)
Revenue Plus	0.065	0.103**	-0.023	-0.016	0.095**	0.039	0.024	0.032
	(0.047)	(0.047)	(0.031)	(0.043)	(0.042)	(0.045)	(0.027)	(0.033)
Flexible Bonus	0.025	0.060	-0.052	0.017	0.040	0.022	0.004	0.014
	(0.035)	(0.046)	(0.040)	(0.039)	(0.041)	(0.046)	(0.032)	(0.026)
Panel B: Current								
Revenue	0.105***	0.077**	0.014	0.014	0.164***	0.099**	0.030	0.035
	(0.036)	(0.037)	(0.016)	(0.029)	(0.045)	(0.045)	(0.027)	(0.023)
Revenue Plus	0.078	0.101*	0.011	-0.035	0.083*	0.058	0.013	0.011
	(0.051)	(0.052)	(0.018)	(0.042)	(0.046)	(0.045)	(0.022)	(0.026)
Flexible Bonus	0.013	0.050	-0.023	-0.016	0.042	0.017	0.020	0.005
	(0.036)	(0.032)	(0.026)	(0.034)	(0.042)	(0.037)	(0.021)	(0.019)
Panel C: Arrears								
Revenue	0.152*	0.142	-0.067	0.074	0.029	-0.112	0.037	0.104**
	(0.088)	(0.088)	(0.063)	(0.049)	(0.134)	(0.142)	(0.070)	(0.050)
Revenue Plus	0.019	0.072	-0.015	-0.043	0.161	0.025	0.116*	0.021
	(0.105)	(0.118)	(0.066)	(0.062)	(0.112)	(0.121)	(0.065)	(0.069)
Flexible Bonus	0.132	0.154	-0.087	0.078	0.099	0.109	-0.029	0.020
	(0.097)	(0.111)	(0.071)	(0.075)	(0.107)	(0.127)	(0.070)	(0.058)
N (total)	473	470	470	473	474	474	474	474
Mean of control group	15.671	16.109	-0.187	-0.244	15.722	16.120	-0.214	-0.183

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table A.5: Selection Effects in Reassessment

	Self-reported Tax Payment	Bribe Payment	Frequency of Bribe Payment	Perception of Corruption	Total payment to tax dept. (Tax + Bribe)
Panel A: Commercial					
Re-assessed * Treatment	4190	-577.6	-.1237	.0069	5567
	(3215)	(487.6)	(.1176)	(.0308)	(4757)
Re-assessed	-421.1	-89.23	.015	-.0331**	-286.9
	(2087)	(237.3)	(.0528)	(.0146)	(3314)
Panel B: Residential					
Re-assessed * Treatment	290	-470	-.1959	.0089	-808.4
	(1494)	(385.2)	(.169)	(.0404)	(1911)
Re-assessed	4043***	-261.9*	.0443	-.0075	4185***
	(841.3)	(137.9)	(.098)	(.0208)	(1113)
Panel C: Mixed					
Re-assessed * Treatment	3458	-467.1	-.2783**	-.0057	4010
	(2635)	(529.6)	(.1405)	(.0387)	(3688)
Re-assessed	3435***	-228.9	-.014	-.0138	3552***
	(634.3)	(279.2)	(.0554)	(.0185)	(770.4)
Sample	Full	Phase 1	Phase 1	Phase 1	Phase 1
N - Commercial	4266	2417	2401	2423	2416
Mean of control group - C	6801	1973	.7456	.6543	9505
N - Residential	4118	2677	2668	2693	2677
Mean of control group - R	2660	1439	.5577	.6419	3607
N - Mixed	3104	1817	1810	1827	1817
Mean of control group - M	4058	2116	.8188	.6364	5229

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table A.6: Who gets re-assessed, split by property type

	GARV	Number of floors	Last renovation was ≤ 2 years ago	Land area (sq. feet)	Total covered area, all uses	Main Road	Tax Category	Percent of property commercial	Percent of property commercial and rented	Tax Liability
Panel A: Residential Properties										
Reassess * Treatment	-10295.205	-0.006	0.008	-1155.315	-81.884	-0.204***	-0.220*	-0.057**	-0.008	-1874.704
	(7615.323)	(0.066)	(0.031)	(2049.895)	(291.742)	(0.076)	(0.115)	(0.024)	(0.015)	(1607.862)
Reassess	20543.107***	0.093***	0.070***	697.581	38.233	0.103***	0.151**	0.059***	0.024**	4157.270***
	(5403.263)	(0.034)	(0.016)	(1597.511)	(162.639)	(0.036)	(0.063)	(0.015)	(0.009)	(1135.561)
Panel B: Commercial Properties										
Reassess * Treatment	57957.114*	0.137	0.002	-287.313	1268.465	0.071	-0.127	-0.033	0.133**	10613.316
	(31898.464)	(0.090)	(0.031)	(512.160)	(1368.006)	(0.065)	(0.135)	(0.027)	(0.055)	(6780.352)
Reassess	6172.937	0.102**	0.091***	525.429*	357.418	-0.051	0.168**	0.032**	0.123***	2598.309
	(11062.100)	(0.048)	(0.016)	(289.437)	(226.146)	(0.035)	(0.066)	(0.013)	(0.028)	(2233.140)
Panel C: Mixed Properties										
Reassess * Treatment	35746.005	-0.116	-0.033	427.943	2091.731	-0.019	-0.125	-0.020	0.031	8462.756
	(30682.690)	(0.091)	(0.041)	(290.318)	(2111.538)	(0.081)	(0.121)	(0.030)	(0.029)	(6883.216)
Reassess	-10220.960	0.199***	0.140***	-159.306	-1256.762	0.016	0.188***	0.071***	0.062***	-2103.943
	(21804.635)	(0.046)	(0.019)	(137.351)	(1533.210)	(0.042)	(0.054)	(0.015)	(0.015)	(4869.901)
N - Residential	5057	5398	5326	5398	5397	5398	5057	5383	5383	5057
N - Commercial	4631	4796	4726	4796	4791	4796	4631	4738	4738	4631
N - Mixed	3342	3523	3475	3523	3523	3523	3342	3495	3496	3342
Mean of control - R	17196.900	1.625	0.033	3293.352	2956.229	0.340	3.850	0.021	0.008	2968.710
Mean of control - C	64919.095	1.418	0.052	1804.940	1877.166	0.602	3.900	0.936	0.503	12487.703
Mean of control - M	47771.270	1.858	0.060	1808.915	2823.112	0.520	3.548	0.334	0.194	8090.962

Table A.7: Impacts on Bribes and Tax Payments, Sub-treatments

	Self-reported Tax Payment	Bribe Payment	Frequency of Bribe Payment	Perception of Corruption	Total payment to tax dept. (Tax + Bribe)
Panel A: Random Sample					
Revenue	562	1649	.2	-.0348	952
	(674)	(1122)	(.161)	(.0374)	(1100)
Revenue Plus	-888***	615	.113	.0258	-1246
	(334)	(777)	(.128)	(.0363)	(924)
Flexible Bonus	-.433	1528**	.264*	.0581*	1059
	(294)	(643)	(.142)	(.0338)	(727)
N	9632	6021	4802	6050	5604
Mean of control group	4428.5	3461.5	0.8	0.6	7616.7
Rev. vs. Multitasking p.	0.134	0.605	0.947	0.023	0.327
Objective vs. Subjective p.	0.697	0.624	0.492	0.038	0.154
Equality of Schemes	0.013	0.520	0.657	0.044	0.048
Joint significance	0.019	0.092	0.214	0.078	0.090
Panel B: Re-assessed					
Re-assessed * Revenue	867	-1238	-.0872	-.0151	-2343
	(2008)	(1117)	(.177)	(.0353)	(2838)
Re-assessed * Revenue Plus	2539	-514	-.191	.012	3742*
	(1667)	(730)	(.128)	(.0323)	(2256)
Re-assessed * Flexible Bonus	3324*	-1039	-.14	-.00587	3630
	(1895)	(735)	(.127)	(.0334)	(2745)
Re-assessed	3369***	-18.6	.00399	-.0195	3792***
	(844)	(234)	(.0435)	(.0131)	(1268)
N	13693	8235	6993	8268	8231
Sample	Full	Phase 1	Phase 1	Phase 1	Phase 1
Mean of control group in random sample	5358.5	2984.8	0.7	0.6	8356.4
Rev. vs. Multitasking p.	0.295	0.691	0.674	0.619	0.024
Objective vs. Subjective p.	0.388	0.857	0.997	0.903	0.263
Equality of Schemes	0.554	0.787	0.871	0.795	0.073
Joint significance	0.255	0.401	0.409	0.926	0.085

Robust standard errors, clustered by circle, in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table A.8: Impacts on Nonrevenue Outcomes, Sub-treatments

	(1) Quality	(2) Satisfaction	(3) Inaccuracy	(4) Tax Gap
Panel A: Main Treatment				
Re-assessed * Treatment	0.008	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)
Panel B: Subtreatments				
Re-assessed * Revenue	0.001	0.001	0.001	-0.012
	(0.032)	(0.032)	(0.025)	(0.039)
Re-assessed * Revenue Plus	0.019	0.019	-0.029	0.014
	(0.032)	(0.032)	(0.021)	(0.040)
Re-assessed * Flexible Bonus	0.031	0.031	0.019	-0.008
	(0.032)	(0.032)	(0.024)	(0.037)
Re-assessed	0.046***	0.046***	-0.060***	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 random sample	0.538	0.555	0.339	-0.103

Robust standard errors, clustered by circle, in parentheses
* 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, 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

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

Several observations are worth making about the condition in (A.2). 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.2) 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.2) reduces to

$$1 - \alpha f'(0) - \beta g'(0) > r \tag{A.3}$$

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 genreal cost 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 if $f'' > 0$ or $g' > 0$ could be at an interior

level of τ.

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.1) 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 precipiously, 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 optimium 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 tax payer is $-\tau - b - \alpha b$ and the utility fo 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

$$\max_b \{r\tau + (1-\beta)\, b\} \{-\tau - (1-\alpha)\, b\}$$

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 parametrers at which collusion takes places is decreasing in r, α, and β, and conditional on collusion taking place, bribes are increasing in r.

B Training and Assessment of Inspectors' Knowledge of Treatments

A total of five training sessions for circle staff were conducted over the treatment period; three in the first year (August, February, May 11-12) and two in the second (August, October, 12-13). Trainings were conducted after each randomization lottery, and retraining sessions were conducted after an interval of 2 - 4 months, covering all treatment staff. The training for each subtreatment was conducted separately to avoid any confusion between them. During the training session treatment staff was given detailed description of each subtreatment, the variables on which their performance would be judged and the formula for calculating individual payments. Circle staff worked through examples calculating their earnings under different scenarios and the sessions concluded with a question and answer session.

To judge circle staff's understanding of the treatments they were quizzed six times at regular intervals over two years. The quiz tested staff on knowledge of the subtreatment they were selected in, the criteria they would be judged on, the formula for calculating their payments and ability to calculate their own payments under different scenarios. The initial understanding of the treatments was low, with around 37% inspectors failing the quiz conducted after the first training, but this improved substantially after multiple trainings, and within 7 months after randomization failure rates dropped to under 1% and remained under 10% for the remainder of the treatment period. The quiz results show that shortly into the project, staff had a good understanding of their respective

subtreatments, how their payments would be calculated and what they needed to do to increase their payments.

Finally, to check that inspectors in fact knew what schemes they were in, towards the end of the second year (in July) all inspectors (including those not in treatments) were called in a brief phone survey. As part of this, they were asked whether they were participating in one of the circle-level incentives schemes, and if so, which one. Appendix Table A.3 reports the result of a dummy for being in each of the 4 inspector-level schemes (Revenue, Revenue Plus, Flexible Bonus, or Information) on actual participation (instrumented with the results of the randomization). As is evident from the table, inspectors clearly both knew which scheme they were in and were able to accurately differentiate between the schemes, with the only mistake being that a small number of inspectors from the Revenue scheme mistakenly reported being in the Revenue Plus scheme.