Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings

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

July 6, 2018

Abstract

Bureaucracies often post staff to better or worse locations, ostensibly to provide incentives. Yet we know little about whether this works, with heterogeneity in preferences over postings impacting effectiveness. We propose a performance-ranked serial dictatorship mechanism, whereby bureaucrats sequentially choose desired locations in order of performance. We evaluate this using a two-year field experiment with 525 property tax inspectors in Pakistan. The mechanism increases annual tax revenue growth by 30-41 percent. Inspectors that our model predicts face high equilibrium incentives under the scheme indeed increase performance more. Our results highlight the potential of periodic merit-based postings in enhancing bureaucratic performance.

*This project is the result of collaboration among many people. We thank Robert Gibbons, Parag Pathak, and numerous seminar participants for helpful comments. We thank Gabriel Kreindler, Alyssa Lawther, Ismail Khan, Kunal Mangal, Guillermo Palacios Diaz, Wayne Sandholtz, Garima Sharma, and Mahvish Shaukat for outstanding research assistance in Cambridge and Zahir Ali, Osman Haq, Turab Hassan, Obeid Rahman, Ali Abbas, Adeel Shafqat, Omer Qasim, Mehak Siddiquei, Shahan Shahid and the late Sadaqat Shah for outstanding research assistance in Lahore. We thank all the Secretaries, Director Generals, Directors, the two Project Directors from the Punjab Department of Excise and Taxation, the Punjab Finance, Planning and Development departments and the Chief Secretary and Chief Minister's offices for their support over the many years of this project. This RCT was registered in the American Economic Association Registry for randomized control trials under trial number AEARCTR-0002753. Financial support for the project came from 3ie, the IGC, and the NSF (under grant SES-1124134). 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

Governments face many constraints in providing incentives to workers. Pay is often subject to strict civil service regulations that make it a mechanical function of tenure and education, with little room for rewarding performance. Scope for promotion can also be limited and mechanical, often restricted by legally-protected seniority-based promotion systems, limiting the potential of career concerns (a la Gibbons and Murphy 1992; Holmström 1999) to provide incentives as well.

In such rigid bureaucratic environments, one way in which managers can provide incentives is through their control of where people are posted. One often hears anecdotal stories of bad performers in a tax or police bureaucracy being sent to remote and unattractive locations as a punishment, or good performers being sent to an attractive location as a reward for exemplary performance. And indeed, in many bureaucracies, transfers occur quite frequently. Yet despite the potential for postings to be used as an incentive device, in practice, in most bureaucracies, many other factors other than performance – such as personal or political connections, or idiosyncratic preferences of managers, or simply bureaucratic arbitrariness – are used to assign positions (Iyer and Mani, 2012). And even to the extent that performance matters for postings, the ambiguity of assignment rules in most contexts limits the degree to which they provide *ex ante* incentives to improve performance.

Even if one wanted to use postings to provide incentives, doing so in practice is complicated as workers may have heterogeneous preferences: what is a desirable posting for one person may be a terrible posting for another. Therefore, to use postings as an incentive, managers need to take these diverse preferences into account. An additional challenge that arises in doing so is the problem of preference revelation: the manager must get the workers to truthfully reveal their preferences, knowing that those preferences will be used to provide them with incentives. And even if the incentive aspects of such a system can be worked out, the allocation of people to tasks induced by such a system, and the disruption caused by moving people around, may cost the organization more than the performance gains induced by the incentive effects.

These challenges aside, such mechanisms can be extremely attractive both because they impose a lower fiscal burden on the state, and also because they provide a measure of fairness and merit in an otherwise fairly regular phenomenon of transfers. Moreover, unlike performance pay, which is less effective in environments with high rents, transfer mechanisms can actually exploit the existence of (future) rents in order to obtain more desirable present behavior (similar to re-election motives inducing desirable political behavior).

In this paper, we ask whether one can explicitly leverage the ability to transfer as an incentive device through a large-scale field experiment in a real-world government bureaucracy: the property tax department of the Province of Punjab, Pakistan. We randomize entire groups of tax inspectors into a system where postings will be explicitly based on performance (and where they know this in advance), or into a control group where postings operate as in the status quo.

In order to do so, we designed and implemented a transfer mechanism, which we term a *performance-ranked serial dictatorship* (henceforth, PRSD). We build on the theoretical litera-

ture on allocation problems (e.g. Abdulkadirŏlu and Sönmez 1998; Svensson 1999), which shows that serial dictatorships – where individuals are ordered somehow, and take turns choosing among remaining positions – are the unique strategy-proof mechanism (in the sense of inducing individuals to report their true preferences) for efficiently allocating a fixed set of slots to individuals when individuals' preferences are unknown. This literature is often silent, however, on how the individuals should be ordered, and typically they are ordered randomly, to create what is known as a "random serial dictatorship".

Our mechanism uses the ordering of individuals in a serial dictatorship to provide incentives. Specifically, individuals are ordered based on their performance. The highest performing individual gets first choice of posting, the second-highest performing individual gets second choice, and so on. Individuals' incentives come through the fact that as they increase their performance, they increase their likelihood of getting a higher position in the serial dictatorship, and thus possibly a more preferable slot. Since performance-ranked serial dictatorships are a special instance of a serial dictatorship, they are also strategy-proof, in the sense that revealing one's true preferences over slots is always a dominant strategy. The idea of creating the ordering in a serial dictatorship based on merit has been used before: the Army, for example, creates its priority list for allocating new offers to specific services based on an "order of merit," a weighted average of academic performance, physical fitness, and military performance (Sï¿œnmez, 2013); what is new in this context is that the primary purpose of the mechanism, and in our study of it, is the *ex ante* incentives to improve performance the system creates based on the desire to move to higher positions in the list.

The incentives to increase effort embodied in such a mechanism are complex and heterogeneous across individuals. In particular, the incentive effects depend on an individual's own preferences among slots, the preferences of others, the expected distribution of everyone's performance outcomes, and differences in the elasticity of effort to motivation. For example, if an individual i's most preferred slot j is ranked very low by everyone else, individual i faces weak incentives, since he will get j with very high probability regardless of his effort. On the other hand, even if individual i is the only person who ranked slot j as a first choice, if slot j is highly ranked by others, individual i still needs to exert effort to ensure that his slot is not taken by another individual who doesn't get his first choice and who prefers slot j to his other remaining options. Analogously, if individual i anticipates being very likely or very unlikely to outperform others in their comparison group due to factors exogenous to their effort, he is likely to have lower *ex ante* incentives.

We begin by setting out a simple model that describes the incentives created by the performance-ranked serial dictatorship as a function of preferences and expected outcomes. We then simulate the equilibrium effort under the model using rich data on the complete set of preferences over postings that we elicited at baseline from all of our tax inspectors, as well as predicted performance under the status quo. This allows to characterize the heterogeneity in incentives across tax inspectors that would be induced by the scheme based on their preferences, under differing assumptions about how much inspectors know about the preferences of others and the degree to which certain individuals expect to perform better than others.

We then analyze the impact of this type of lateral allocation scheme through a randomized field experiment we conducted over two years in a real bureaucracy. We worked with the Provincial Excise and Taxation Department of Punjab, Pakistan, which is comprised of approximately 500 tax units, or "circles." Each circle covers a predefined geographic area, and is staffed by an "inspector." Within each district (i.e. roughly the same metropolitan area / county), we randomly assigned circles into groups of about ten circles each. There was substantial heterogeneity within groups in circle characteristics – for example, even within districts, the 90th percentile circle has a tax base more than three times as large as the 10th percentile circle. An examination of our preference data suggests that while there are clearly more "popular" circles that many inspectors like, there is also a substantial idiosyncrasy in preferences, including a substantial status quo preference, such that inspectors face (predictably) differential incentives under the scheme.

The experiment took place as follows. At the beginning of the first year, inspectors in a randomly-selected half of the groups were told that at the end of the first year their job postings within the group would be reassigned using a performance-ranked serial dictatorship, where the ranking was done based on the year-on-year improvement in a metric of circle level tax performance.[1] At the end of the year, postings within the group were re-assigned based on preferences, as per the mechanism. Control group postings continued under business-as-usual. Groups were re-randomized again at the beginning of Year 2 into treatment and control groups (stratified based on the first year treatment status), and again treatment groups were told that postings would be allocated based on a performance-ranked serial dictatorship. The transfers were implemented again as promised for second year treatment groups at the end of Year 2.

We find that overall, the promise of performance-based postings substantially raised revenues. We estimate that revenues were about 5 log points higher in treatment groups than in control groups in the first year, and 9 log points higher in treatment groups than in control groups in the second year.[2] This amounts to an increase in the growth rate of tax revenues of 41 percent in the first year and of 30 percent in the second year. Note that this is a pure incentive effect – this is the effect on revenue merely from announcing that the scheme will be applied to determine postings in the subsequent year.

These magnitudes are substantial not just in an absolute sense, but also compared to a performance pay scheme we evaluated in the same context. In particular, a previous randomized trial we conducted in a largely overlapping set of locations in the years prior to this experiment (2011-2013) showed that on average, piece rate schemes paying the tax inspectors an average of 10 percent of every marginal dollar collected led to increased tax revenues by about 9 log points (Khan, Khwaja

[1] In half the groups, the metric was randomly selected to be year-on-year growth in actual tax revenue; in the other half, the metric was year-on-year growth in tax assessments.

[2] Groups performed broadly similarly regardless of whether they were incentivized based on revenue or tax assessments, though the estimates suggest revenue-based estimates were more effective in the second year. Note that after the design for this project was finalized, analysis we conducted in Khan, Khwaja and Olken (2016) showed that the main mechanism for raising revenue in that study was actually increasing tax assessments. Given this, it is not surprising that the two performance metrics used – revenue and size of tax base – end up being largely similar in this context (See Appendix Table A.2).

and Olken, 2016), with the most successful piece-rate scheme achieving a 13 log point increase. This means that the increase in tax revenues from the performance-based postings scheme in the second year had an impact between two-thirds and equal the size of very substantial piece-rate schemes – but at zero fiscal cost to the government.

We then take the theory to the data to see whether or not those workers whom the model predicts should face stronger incentives under the performance-based serial dictatorship do, in fact, respond more to the scheme. In particular, we simulate the equilibrium effort implied by the PRSD under the model under varying assumptions for two factors that determine the strengths of these incentives: (i) whether individuals know the preferences of others and (ii) whether they can can forecast their likely place in the performance distribution under business-as-usual.

We find evidence that those individuals who were predicted by the model to have stronger incentives under the scheme indeed increase their revenue more in response to being randomly allocated to the performance ranked serial dictatorship scheme. These results imply that collecting information on preferences for slots and the expected distribution of outcomes can allow one to reliably predict in which contexts such incentive schemes have the greatest potential to improve performance. More generally, this also builds on the small empirical literature, largely (though not entirely) drawing from sports, on how tournament-based compensation schemes can lead to heterogeneous effort choices depending on individuals' forecast of the likely incentives they face from the tournament (e.g., Prendergast, 1999; Sunde, 2009; Boudreau, Lakhani and Menietti, 2016), and shows how these ideas can be applied in a case where the heterogeneity in the prize (in our case, a job posting) comes from an allocation decision.

While our results highlight the incentive properties of a performance-based serial dictatorship, a related question is the allocation of people to slots it induces. Where individuals finally end up will be determined by the composition of workers' preferences – and in particular, by what the highest performing workers prefer, since their preferences will get implemented with high probability. In our context, it turns out that high performers tend to largely prefer larger circles. Therefore, on net, treatment areas end up allocating the highest performers to larger circles than in control areas. In our context, if we think that the effect of workers on tax revenue is proportional (that is a high performer can increase tax revenue more in a larger circle), this allocation may serve to further enhance the impacts of the scheme.

These effects thus far have focused only on the first year the program was announced – when the program had incentive effects, but the allocation effects had not been realized. By randomizing treatments again in the second year, we can also examine dynamic effects of the program. We find that the effects on revenue persist in the second year even for those inspectors no longer exposed to incentives, but after the postings from the scheme had been implemented. One reason for this may be that, in this context, the prime mechanism to increase tax revenues is to add new properties to the tax rolls; once added, they continue to pay taxes. Allocation effects induced by the scheme could also contribute; indeed the point estimates are larger in the second year, after the incentives are no longer in place, than they were initially. While we cannot definitely disentangle mechanical

persistence from allocation effects, the key point is that the allocation effects do not appear to be negative.

We include one cautionary note, however: the results show that exposing an inspector to the scheme for two years in a row has no larger an effect than exposing the inspector to the scheme once, and if anything the point effects are lower. One possible explanation, consistent with the theory, is that over time inspectors learn more information about the preferences of others and their likely place in the outcome distribution. Our simulations suggest that, in general, more information about the preference matrix and the likely outcome distribution tends to diminish incentives; this would be consistent with diminished effects over time as people learn more about the preferences of others and their likely performance outcomes. It could also be that inspectors find being at risk of losing their posting right after they "earned it" discouraging thereby dampening effort for those who included in the scheme a second time.

On net, the results suggest that bureaucracies can use workers' preferences over their allocations to jobs to create incentives. The results suggest that the degree to which this approach will be useful depends on several factors, all of which can be checked by collecting some simple preference data at baseline. First, both the model and the empirical results suggests that the strength of the incentives will depend on (beliefs regarding) the homogeneity of preferences and the distribution of expected outcomes; by collecting baseline data on preferences and business-as-usual performance, one can use the model to simulate whether the incentives are likely to be weak or strong. Second, the degree to which the allocations induced by the scheme are likely to further the principal's objectives depends on what characteristics make a location popular. In our setting, top performers preferred larger circles, suggesting that allocation effects would serve to further increase tax revenue; but in other contexts (for example, assigning teachers to schools), the principal might not want top performers in the spots they desire most. By analyzing what types of characteristics make a location popular, the principal can decide whether allocating top performers to those types of locations is likely to further the principal's overall objectives. Finally, our results also suggest that these schemes - especially to the extent that they generate actions that have persistent effects - may work better when periodically applied.

The remainder of this paper is organized as follows. Section 2 describes the context we are working in, the data and intervention. Section 3 outlines a model that characterizes the marginal incentives implied by the performance-ranked serial dictatorships. Section 4 presents the experimental design and empirical strategy. Section 5 then presents the main empirical findings as well as additional results, including the heterogeneous impacts implied by the model, the allocation impacts of the scheme, and the dynamic effects. Section 6 concludes.

2 Setting, Data, and Intervention

2.1 Setting: Property Tax Administration in Urban Punjab, Pakistan

This study takes place in urban areas of Punjab, Pakistan. Punjab is quite large – its population of 110 million would rank it twelfth in the world were it a country – and the cities that we study here include a number of large metropolitan areas, ranging from over eleven million people in Lahore to about two million in Multan and Gujranwala. Like other low income countries, tax collection in Pakistan is generally low, with issues ranging from a low tax base to inadequate enforcement and corruption.

In this study we work with the Punjab Excise and Taxation department's property tax division. The property tax is not only one of the most significant of regional taxes, but is one where there is substantial returns to effort, and also substantial potential for tax evasion. Punjab's urban property tax is computed based on a formula that takes into account the square footage of land and buildings on the property, multiplied by standardized values from a table that depends on neighborhood wealth status, residential, commercial or industrial property status, whether the property is owner occupied or rented, and location (i.e. on or off a main road).

The primary unit of tax collection is the "circle," a predefined geographical area that covers anywhere from two to ten thousand unique properties. The circle is led by an inspector who is essentially responsible for all stages of the tax: he determines each property's tax liability, sends an annual tax bill to the property owner, and is ultimately responsible for collections and dealing with any issues raised by the taxpayer. Together with a clerk in charge of record keeping and a constable who assists the inspector in the field, the team maintains a record of all properties and their attributes (size, type of use, etc.), apply the valuation tables to each property, and determine which exemptions apply.

Although property tax should be formulaic, property tax inspectors play a key role in tax administration, because they are the only source the government has for the inputs into this formula, how the formula is applied, and for even discovering which properties exist in the first place and should be taxed. Not surprisingly, collusion between taxpayer and tax inspectors is thought to be widespread, reducing government revenue.

As is common for civil servants in developing economies, tax officials receive fairly low wages that are rarely, if ever, tied to performance. In our previous study (Khan, Khwaja and Olken, 2016), we showed that tax inspectors respond to performance pay – offering the three tax officers (inspector, constable, and clerk) performance pay equal to a total of 30 percent of all taxes collected above a historical benchmark increases taxes by 9 log points. The additional revenue comes about not by overtaxing, but rather by adding new properties to the tax rolls and by eliminating undertaxation. The current study takes place in essentially the same setting, with sufficient overlap so that the comparison between the two is meaningful.[3]

[3]Due to our desire to create proximate groups of 10 or so circles within which the PRSD scheme was applied (as we describe in more detail below), the current sample is somewhat smaller, with around 81% of the tax-circles in our prior study part of the current paper sample. We should note though that in terms of inspectors only 3% of the

This paper instead focuses on an alternative incentive mechanism – the assignment of inspectors to circles. Such postings – either to better or worse assignments – are often the primary tool available to supervisors who want to improve performance. In our context, in fact, these transfers of inspectors between circle are fairly common, with about one-third of tax inspectors typically transferred between circles each year. Inspectors are likely to care about where they are posted, as there is substantial heterogeneity across circles. For example, the number of properties varies substantially – the 90th percentile circle has more than three times as many taxable properties as the 10th percentile circle. Even more important is heterogeneity in ease of collecting taxes, opportunities for corruption, and amenities – all of which can be used to provide incentives. However, in practice, the transfer process is opaque and subject to political influence, so their use as an incentive device is in practice limited. The fact that political influence and other factors other than performance often influence postings is common in many settings, particularly for those outside the very top of the civil service (see, e.g. Iyer and Mani 2012).

2.2 Data and Summary Statistics

Our primary data source is circle-level administrative data on tax performance. The administrative data is based on the quarterly reports that each inspector files, which show their overall collections (separately for current year and past years/arrears collections) and the total assessed tax base. We digitized these reports for all tax circles.[4]

Summary statistics for key variables from the administrative data are shown in Table 1 for the second year of the experiment (FY 2015). First, current year tax revenues are substantially larger than arrears (i.e. collections against past years' unpaid taxes) – the mean of log current year tax revenues is 16.00 compared with just 13.54 for log arrears, implying that, on average, current revenue is about twelve times as large as arrears. This suggests that the main impacts on total revenue will likely be felt through increases in current year revenue. Second, the log recovery rate (the log of tax revenue divided by the tax base net of exemptions) is -0.08 for current year taxes, which implies that about 92 percent of all taxes that are demanded by the government are in fact paid. In addition to nonpayment, a substantial issue is that many properties are either underassessed or not assessed at all. Tax inspectors can therefore respond to performance incentives by adding new properties to the tax rolls, assessing existing properties more accurately, and increasing collections of existing assessments.

In addition, we also collected rank-order baseline preference data from all inspectors over all circles in their (randomly assigned) groups, which consisted of an average of ten circles from within the same metropolitan area (more details on the construction of these groups below). Inspectors

inspectors were treated under the interventions introduced in both these papers (for all four years), with 62% having been treated at least one. Thus, while there is sufficient overlap to draw meaningful comparisons between the two studies, it is not the case that the inspectors we constantly under some experimental scheme for four straight years.

[4]In our previous paper (Khan, Khwaja and Olken, 2016) we also selected a random sample to be verified each year by aggregating (thousands of) bank-verified receipts of individual payments; we found no statistically or economically significant discrepancy between the administrative data and our independent verification.

were given a preference form prior to the assignment of treatment status, and were told to rank all circles in their group from 1 to J, with 1 as the highest ranked circle.[5]

Before beginning our analysis of incentive effects, it is useful to examine the preference data a bit further. We find that preferences have both a common and idiosyncratic component; in particular, many inspectors seem to display a strong preference for their status quo allocation. To see this, Figure 1a shows the distribution of inspectors' ranks of their current position at baseline. We normalize ranks such that 1 is the highest rank and 0 is the lowest rank. Figure 1a shows that about half – 53 percent – of inspectors rank their own circle as their most preferred, with the rest expressing a desire to move.

To examine the common component of preferences, for each pair of inspectors within the same group, we calculate the pairwise correlation between their ranks. Figure 1b plots the distribution of these pairwise correlations, along with what the distribution would look like if preferences were randomly distributed. As is evident from Figure 1b, the distribution of correlations is substantially shifted to the right compared with one would expect from random chance; the mean correlation between inspectors' ranks is about 0.22.

Given this preference data, it is also useful to estimate the degree to which the current allocation is Pareto inefficient, from the perspective of maximizing inspectors' utility. Note that any allocation that results from a serial dictatorship will always be Pareto efficient in this sense, so to the degree the current allocation is far from the Pareto frontier, there may be large gains in inspector utility from implementing the scheme even holding effort constant. One way to characterize this is to calculate the core allocation of inspectors to circles, using Gale's Top Trading Cycle algorithm (Shapley and Scarf, 1974). This algorithm computes the unique allocation of inspectors to circles such that no inspector is worse off than he is in the status quo, and no inspector or group of inspectors would want to deviate. The difference between the status quo and the core is a measure of how inefficient the current allocation is. We find that 15 percent of inspectors would be able to move to a posting they strictly prefer to the status quo in the core. Conditional on moving, these individuals move to circles ranked about 30 percentiles higher in their preference ordering. The relatively small number of movements suggests that while there is some Pareto-improving room for improvement on the status quo, it is limited. The re-allocations induced by the scheme will therefore largely be non-Pareto-improving, in the sense that increases in utility for some are likely to lead to decreases in utility for others. We will return to this when we consider heterogeneous impact of the scheme.

2.3 The Performance-Ranked Serial Dictatorship Scheme

We now describe the basic design of the Performance-Ranked Serial Dictatorship Scheme that was introduced in collaboration with the Excise and Taxation department for a two year period

[5]Inspectors had incentives to reveal their true preferences. The scheme was explained briefly to inspectors, so they could understand that truthful revelation was a dominant strategy, and inspectors were told that if they were chosen for the scheme, these preferences would be used in assignment (though they were, *ex post*, given an opportunity to revise preferences).

beginning in 2013. We describe the theoretical properties of this scheme in the subsequent section.

The primary goal of the scheme was to incentivize performance by linking performance explicitly to postings. The scheme was known formally within the Excise and Taxation department as the "Merit-Based Transfers and Postings" (MBTP) scheme to make this link clear.

The scheme worked as follows. Within each of the ten major metropolitan areas in Punjab, we randomly allocated circles into groups of approximately ten circles each.[6] At the beginning of the tax year (i.e. in July), groups were randomly selected to either participate in the MBTP scheme or to remain in the status quo.

For groups selected to be in the MBTP scheme, all inspectors were told that they would be ranked based on their performance, and then based on this ranking, would be given a choice of circles within their group. Specifically, inspectors were told that if they were the top-ranked inspectors in a group they would be posted in their first-choice circle, the next ranked inspectors would be posted in their top preference from the remaining circles, and so on.[7] Performance was calculated in two ways (randomized by group): for one sub-treatment (the "recovery" group), inspectors' performance was calculated as the year-on-year percent increase in their current circles' tax collected during a fiscal year; for the other sub-treatment (the "demand" group), inspectors' performance was calculated as the year-on-year percent increase in their current circles' assessed tax base.

The scheme was implemented as promised. At the end of the fiscal year (but before final performance had been calculated), inspectors submitted their final, binding set of preferences over all circles in their group.[8] Postings were then carried out as described: the top-rank inspector was given his first choice of posting, the second-rank inspector was given his top choice among remaining circles, and so on.

Note that postings were done within groups, which as described above were randomly selected groups of ten circles within metropolitan areas. The fact that the scheme was done within metropolitan areas ensured that no inspector would need to physically move his family as a result of the scheme. The fact that choice was constrained to groups of ten circles was for experimental feasibility, so that there would be both treatment and control areas within each metropolitan area. While ten circles still entails substantial heterogeneity – within groups, the 90th percentile circle has tax revenue almost nine times larger than the 10th percentile circle – the incentive effects we find here are most likely an underestimate of the incentive effects that would be generated if choice

[6]The major metropolitan areas correspond to "divisions" in the tax department, with the exception of the capital city of Lahore. Lahore consists of two divisions, but they were combined to form a single metropolitan area for the purposes of forming groups.

[7]In order to convince inspectors in the first year that postings would be made as promised, two additional groups of ten inspectors each were randomly selected to have the merit-based postings implemented at the start of Year 1, based on performance in the previous year following the same PRSD scheme. We exclude these twenty inspectors from the analysis here.

[8]Although final postings were made on the basis of total performance during the fiscal year, inspectors were given information at the end of the third quarter as to their tentative ranking before submitting their final preferences. The final preferences were similar, but not identical, to the baseline preferences: the average rank-correlation of between initial and final preferences is 0.63.

was given over a larger number of locations.

3 Modeling Incentives under PRSD

The incentives created by the performance-ranked serial dictatorship are complex, and depend on the relationship of an inspector's preferences with those of everyone else and how he expects his performance to compare to others. This section describes a simple model to characterize these incentives, and then simulates the model using preference data we collected at baseline to better understand the heterogeneity in incentives under the scheme.

3.1 Allocation under the PRSD Scheme

Suppose that inspector i obtains utility u_{ij} from being assigned to circle j. This determines a preference ordering over circles for each inspector i. We denote the overall preference matrix implied by these preferences from all inspectors by $\mathbf{P}$**.** Further, suppose that the outcome (in our case, growth in tax revenue) for inspector i is given by

$$y_i = y_{i0} + e_i + \epsilon_i \tag{1}$$

where e_i is the effort from inspector i, y_{i0} is the growth rate that would be observed in the absence of effort (which may differ across circles), and ϵ_i is an iid error term with standard deviation σ_ϵ.[9]

For any vector of outcomes $\mathbf{y}$, the PRSD allocation mechanism, combined with the preference matrix $\mathbf{P}$, yields an allocation $r_i(\mathbf{y}, \mathbf{P})$; that is, for a given preference matrix $\mathbf{P}$, any realization of outcomes $\mathbf{y}$ yields a mapping of assignments of inspectors to new circles given by $r_i(\mathbf{y}, \mathbf{P})$, defined such that inspector i is allocated to circle j if $j = r_i(\mathbf{y}, \mathbf{P})$**.** The function $r_i(\mathbf{y}, \mathbf{P})$ implements the serial dictatorship given preferences $\mathbf{P}$ and the ordering from $\mathbf{y}$; that is, for a given group g (we suppress group classifiers for notational simplicity) the inspector i with the highest y_i is given his first choice circle amongst the set of circles in group g, the inspector i with the second highest y_i is given his first choice from among all remaining circles, and so on.

Suppose that the cost of effort for inspector i is given by the convex function $c(e_i)$. Then inspector i will choose effort to maximize his expected utility:

$$\max_{e_i} \sum_{\mathbf{j=1}}^{\mathbf{J}} u_{ij} Pr(j = r_i(\mathbf{y}, \mathbf{P})) - c(e_i) \tag{2}$$

In solving this expression, inspector i takes the effort from other inspectors as given. We can

[9] Note that one might also be interested in the impact of incentives on the bargaining relationship between taxpayer and tax inspectors, which we study in detail in Khan, Khwaja and Olken (2016), and which is not captured by (1). In this paper, we take as given that the principal is interested in incentivizing tax collection. Since we are interested primarily in exploring the properties of a performance-based allocation scheme, in this paper we therefore abstract from the details of this bargaining game, and instead model the response to incentives using a more traditional moral hazard framework with unobservable effort.

therefore rewrite this as

$$\max_{e_i} \sum_{\mathbf{j=1}}^{\mathbf{J}} u_{ij} Pr(j = r_i(y_i, \mathbf{y_{-i}}, \mathbf{P})) - c(e_i) \tag{3}$$

The first-order condition governing effort for each inspector i is given by

$$\frac{dE\left[u\right]}{de_i} = \sum_{\mathbf{j=1}}^{\mathbf{J}} u_{ij} \frac{\partial Pr(j = r_i(y_i, \mathbf{y_{-i}}, \mathbf{P}))}{\partial y_i} = c'(e_i) \tag{4}$$

An equilibrium in the model is a choice vector $\boldsymbol{e}$ such that equation (4) is satisfied for all inspectors simultaneously.[10]

To build intuition, it is worth noting that the first-order condition suggests that there are several factors one would expect would influence the effort decision of a particular inspector i. The first factor is the preference matrix $\mathbf{P}$. If all inspectors i have identical preferences, then moving inspector i's outcome y_i up one rank in the y distribution moves inspector i up one rank in his preference distribution. To simplify notation, label the circles j such that 1 is the lowest-ranked circle and J is the top-ranked circle.[11] The FOC in this case can then be simplified to be

$$\sum_{\mathbf{j=1}}^{\mathbf{J}} u_{ij} \frac{\partial Pr(Rank(y_i, y) = j)}{\partial y_i} = c'(e_i) \tag{5}$$

where $Pr(Rank(y_i, y) = j)$ denotes the probability that inspector i is ranked j'th in the distribution. Note that while rank statistics like this are difficult to compute analytically, they can be easily simulated, as we discuss in more detail below.

An alternative extreme is one in which each inspector has completely different preferences, and in particular, each inspector has a unique first choice circle. In this case, the assignment function assigns inspector i to his first choice circle *regardless of performance*. Relatedly, if for any inspector i his most preferred circle is everyone else's least preferred circle, he will be assigned to his most preferred circle regardless of incentives. More generally, between these two extremes, to the extent that an inspector i's most preferred circle(s) are also highly preferred by others, one would expect that inspector will generally face effectively higher returns from the scheme, though the precise incentives depend on the complete structure of preferences, as we will explore in more detail below.

A second factor that influences effort is the distribution of the $\mathbf{y_0}$ – the predictable component of an inspector's performance – in equation (1). Unlike winner-take-all tournaments, which create steep incentives for those who are potentially near the top of the distribution and little incentive for those who have no chance at winning (Prendergast, 1999), this incentive scheme creates incentives

[10] Since strategy spaces are non-empty and compact (inspectors' actions/effort are limited to at most 24 hours per day, creating an upper bound), and since inspectors' utilities are continuous (due to the inclusion of the ϵ_i *iid* error/noise term in equation 1), there exists a mixed-strategy Nash equilibrium of the game (Glicksberg, 1952). However, the equilibrium need not necessarily be unique; we explore this further in our numerical simulations of the model below.

[11] Note that this is slightly different notation from what we use in the empirical exercises, where we normalize ranks to be on a $[0, 1]$ scale, with 1 as the highest rank. We use the $\{1, ..., J\}$ notation in the theory for ease of exposition, but use the continuous measure in the empirics since different groups have different total numbers of circles J.

throughout the distribution, in the same way that optimal tournaments generally feature prizes for all rank-order positions, not just first place (Lazear and Rosen, 1981), but the degree to which there are incentives depends on how close other inspectors are to you. Intuitively, if the $\boldsymbol{y_0}$ are close together, then a relatively small change in e_i is sufficient to produce a change in rank with high probability holding everyone else's e constant. If the $\boldsymbol{y_0}$ are further apart, then a larger change in e is required to produce the same change in expected rank, so one would expect smaller effort in these cases. Figure 2 qualitatively illustrates different potential configurations of the $\mathbf{y_0}$ distribution. In Figure 2a, the $\mathbf{y_0}$ are close together, with uniformly strong incentives; in Figure 2b, the $\mathbf{y_0}$ are further apart, with lower incentives; in Figure 2c, the $\mathbf{y_0}$ are bunched in the middle but further apart in the tails (as they would be if they were normally distributed, for example), producing higher incentives for those in the middle of the distribution and weaker incentives in the tails. The heterogeneity in incentives from the $\mathbf{y_0}$ in turn interacts in complex ways with the heterogeneity in preferences; we therefore use simulations of the model below to compute numerically the heterogeneity in incentives these two forces create.

A final component that influences outcomes is the u_{ij}'s, the utility of different positions to the inspectors. With common preferences and common $\mathbf{y_0}$'s, the classic result from Lazear and Rosen (1981) suggests that there exists a set of u_{ij}'s such that the scheme would replicate the efficient piece rate scheme in terms of inducing socially optimal effort levels. That said, unlike the Lazear and Rosen (1981) case where the tournament creator chooses the prizes arbitrarily, in this case, the u_{ij} are fixed by inspectors' preferences. Given this, in the more general case, with arbitrary u_{ij}'s, as well as heterogeneity in $\mathbf{P}$ and $\mathbf{y_0}$, the incentives from such a scheme will not necessarily be optimal. Comparing the degree to which these provide incentives to an actual piece rate scheme is thus also of interest to see how close, in practice, the incentives here come to a piece rate.[12]

Since the various components – information about alignment or idiosyncrasy of preferences $\mathbf{P}$, predictable performance y_0, the change in utility from moving up or down a rank (i.e. the u_{ij}'s), and the error variance – all interact in equation (4) to produce incentives in complex ways, one cannot easily characterize the heterogeneity in incentives faced by different inspectors analytically. We therefore simulate the model to calculate the incentives faced by people under the scheme.

3.2 Applying the Model to Context

To better understand how the PRSD mechanism operates in our context, we simulate the model given actual preference and (predicted) performance data in our context. We do so under different assumptions about what information inspectors have about others' preferences $\boldsymbol{P}$ and predictable performance $\boldsymbol{y_0}$. The idea of the simulation is to see what the model would predict in terms of the relative strength of incentives across different inspectors i based on the distribution of preferences $\mathbf{P}$ and the predictable component of performance $\boldsymbol{y_0}$. In Section 5.2 below, we then investigate

[12]While there are number of lab experiments along these lines (e.g. Bull, Schotter and Weigelt 1987),there are relatively few empirical studies along these lines. The study of tournaments by Bandiera, Barankay and Rasul (2013) is one prominent example, which also considers endogenous team formation induced by the different incentive systems.

the degree to which inspectors whom the model predicts should face greater incentives under the scheme do in fact respond more when randomized into the experiment.

To operationalize the model, we begin by characterizing marginal return to effort (i.e. $\frac{dE[u]}{de_i}$ from equation (4)) for a given effort vector $\boldsymbol{e}$, and then we solve for the Nash equilibrium vector of efforts. To calculate the marginal return to effort (i.e. $\frac{dE[u]}{de_i}$ from equation (4)), we need to parameterize the utility over slots, u_{ij}, and estimate the distribution of $\boldsymbol{y_0}$. For the utility over slots u_{ij}, the preference data $\boldsymbol{P}$ we have from inspectors is ordinal, i.e. their ranking over slots. To cardinalize it, we parameterize the utility function over different circles (u_{ij}) linearly, with $u_{ij} = 1$ for inspector i's top-ranked circle and $u_{ij} = 0$ for inspector i's lowest ranked circle.[13] In order to obtain $\mathbf{y_0}$, we first regress actual revenue change on two lags of a circle's (log) revenue and tax base for the control group. We then use both the coefficients and residual from this regression to predict $\mathbf{y_0}$ and σ^2_ϵ (the variance of the error in equation (1)). The distribution of estimated $\boldsymbol{y_0}$ is shown in Appendix Figure A.1.

Before we turn to the equilibrium effort, we first illustrate the importance of different information assumptions, using simulations of the model. Figure 3 plots a histogram of the distribution of marginal incentives $\frac{dE[u]}{de_i}$ under the model across inspectors i evaluated at the point $\boldsymbol{e} = \mathbf{0}$, under different assumptions of knowledge. Since these are evaluated at the point $\boldsymbol{e} = \mathbf{0}$; that is, these are the starting marginal returns to effort at the point the scheme is turned on. [14]

Panel A begins by plotting the distribution of $\frac{dE[u]}{de_i}$ under the assumption of full knowledge; that is, that all inspectors know the full distribution of $\boldsymbol{P}$ and $\boldsymbol{y_0}$. The figures shows substantial heterogeneity in marginal returns across inspectors, with a mass of inspectors at 0, facing effectively no marginal return to effort, and some facing a relatively steep marginal return.

The remaining panels of Figure 3 plot the same figure under alternative assumptions of what inspectors know, turning off first knowledge of $\mathbf{P}$, then knowledge of $\mathbf{y_0}$(predicted performance), and then knowledge of both $\boldsymbol{P}$ and $\mathbf{y_0}$. Note that when we turn off knowledge of $\mathbf{P}$, we need to make an alternate assumption for what inspectors i believe about the preferences $\mathbf{P}$ for all other inspectors. We examine two possible assumptions: 1) that inspectors i believe that inspectors $-i$ have random preferences, and 2) that inspectors i believe that all other inspectors $-i$ have the same preferences they do. As discussed above and shown in Figure 1, preferences have both a common and idiosyncratic component, so reality is likely to be somewhere between these two extremes.

Note that less knowledge – either not knowing $\boldsymbol{P}$ or not knowing $\boldsymbol{y_0}$ – leads to a rightward shift in the distribution of expected utility. That is, adding knowledge about either $\boldsymbol{P}$ or $\boldsymbol{y_0}$ seems to dampen incentives for some people. Intuitively, with knowledge of $\boldsymbol{y_0}$, people may now know that their outcome is less responsive to effort, since they may be predicted to be far apart from other inspectors (panels B and C). With knowledge of P, some people know they are likely to get

[13] We also experimented with alternative functional forms, where we make utility quadratic or cubic in the normalized rank. Results are virtually identical; see Appendix Table A.6, which is analogous to Table 4.

[14] These marginal returns to effort at $e = 0$ are also similar to level-1 reasoning (Camerer, Ho and Chong, 2004; Crawford and Iriberri, 2007), in which each individual i calculates their best response to others business-as-usual responses.

a good outcome regardless of how hard they work, dampening their incentives (panel D). As is evident from the figures, in our context there is substantially more heterogeneity in effort due to $\boldsymbol{P}$ than due to $\boldsymbol{y_0}$ – turning off knowledge of $\boldsymbol{y_0}$ makes a noticeable but small difference, whereas turning off knowledge of $\boldsymbol{P}$ entirely and assuming everyone has identical $\boldsymbol{P}$ eliminates most of the heterogeneity in incentives.

If inspectors know neither $\boldsymbol{P}$ nor $\boldsymbol{y_0}$, all inspectors in a given group have the same incentives. When inspectors assume that all inspectors share their preferences, the incentives are at their maximum (panel e). Intuitively, this is because moving up one rank in the outcome distribution always moves the inspector up to one rank higher preferred circle. When inspectors do not know y but assume other inspectors have random preferences (panel f), incentives are dampened somewhat (by approximately half); intuitively, this is because in some random orderings, inspectors outcomes will not depend on their performance (e.g. if they uniquely prefer a given circle that everyone else ranks poorly).[15] These graphs suggest that the scheme is likely to work best in cases where inspectors have similar preferences and where their predicted outcomes in the absence of effort are as similar as possible. We will return to examine this prediction directly in the empirical results below.

We now turn to simulating the equilibrium effort induced by the scheme. While Figure 3 provided a basic sense of how informational assumptions could vary the strength of the incentives, it only showed the distribution of $\frac{dE[u]}{de_i}$ evaluated at $\boldsymbol{e = 0}$. However, to accurately predict how much inspectors will respond to the incentives, we need to solve for a Nash equilibrium in efforts, i.e. for a vector of efforts $\boldsymbol{e}$ such that equation (4) is satisfied simultaneously for all inspectors. This requires additional assumptions. Specifically, in order to do so we need to parameterize the right-hand side of equation (4), i.e. cost function $c(e)$. While there are of course a range of ways of doing so, we take a parsimonious approach and parameterize the cost function $c(e_i)$ as a simple quadratic, i.e. $c(e_i) = \alpha e_i^2$, with α as an unknown cost of effort parameter.[16] We then estimate α using simulated method of moments so that the *average* effort in the model matches the *average* change in effort induced by the experiment. That is, we choose α so that $E[y_i - y_{i0}]$ from the model matches the experimental estimate of the change in total log revenue estimated via equation (6) below.[17] Remaining details of the simulation exercise are presented in Appendix A.1.

In practice, the full Nash equilibrium results are quite similar to the results shown in Figure 3; the correlation between the marginal returns evaluated at $\boldsymbol{e} = 0$ plotted in Figure 3 and the full Nash equilibrium vector of efforts e is 0.89 for the full knowledge case, and even higher for the other less knowledge cases considered.

[15] The reason that the simulated marginal returns are not exactly identical in panel (f) comes from the fact that there are different numbers of circles in different groups.

[16] We also experimented with alternative functional forms for the cost function. In particular, in Appendix Table A.5 below, we present our main out-of-sample heterogeneity tests using simulated efforts from cost functions $c(e_i) = \alpha e_i^{\frac{3}{2}}$ and $c(e_i) = \alpha e_i^3$, respectively. The results are virtually identical to the main results in Table 4 (discussed in more detail below) which are based on $c(e_i) = \alpha e_i^2$.

[17] We use the estimate of α for Year 1; re-estimating α for Year 2 produces a different α but does not qualitatively change the results.

Note that in fitting the model here we are only using a single, average moment – we are not using any heterogeneity across inspectors in response to treatment to estimate the model. While one could allow for a more flexible cost function and use additional empirical moments to estimate all relevant (cost, informational, utility function) parameters, our parsimonious approach allows us to perform "out-of-model" empirical tests. By only using a single moment to estimate α, our predicted effort levels under the model are not guaranteed to fit any particular pattern of heterogeneity in effort levels in the data. Thus, if we find that inspectors who have higher predicted effort level indeed respond more to being randomized into the PRSD, this serves as a test of the basic model. We examine this test empirically in Section 5.2 below.

4 Experimental Design and Estimation

In this section we first describe the overall research design, which uses a randomized controlled trial to examine the impacts of the PRSD. We then present the primary estimating equations to examine both the overall impact of the PRSD schemes as well to examine potential heterogeneity of impact as implied by the theoretical analysis and simulations in the previous section.

4.1 Experimental Design

We study the impact of the performance-ranked serial dictatorship using a randomized controlled trial. In order to do so, at the start of the first year circles were randomly assigned to be groups of 9-11 circles each, within metropolitan area. This results in 41 groups. After soliciting baseline preferences of all inspectors for circles in their assigned group, groups were randomized into treatment and control areas, stratified by metropolitan area. Within treatment areas, half the groups were randomly assigned to have performance judged by year-on-year change in tax recovery, and half were randomly assigned to have performance judged by year-on-year change in tax assessments.[18]

In the second year, groups were randomized again into treatment and control, stratified based on their treatment in the first year. The re-randomization was done prior to inspectors who participated in the first year submitting their final preferences, which means that the preferences (and allocation) of those inspectors reflects the fact that they know whether they will be continuing in the scheme for a second year. This allows us, in Year 2, to explore both a) the differential effects of Year 1 circles having already experienced the PRSD scheme in the past but no longer receiving the incentives effects, b) the effects of participating in the scheme for multiple years in a row, and c) the pure effect of joining the scheme for the first time in Year 2 relative to pure controls.

The overall treatment assignment matrix as of Year 2 is shown in Table 2, and map is shown for several sample districts in Appendix Figure A.4. Note that if a Year 1 group was randomized

[18]Note that in Year 1, performance was judged using what the department calls "net demand," which represents the total taxes assessed after exemptions are taken into account. Given that there is some heterogeneity in the exemption rate across circles, exemptions are included in the performance metric, and circle staff have little control over the exemption rate, in Year 2 in an effort to simplify the performance metric further, performance was judged only using "gross demand" (which is the taxes assessed before exemptions are taken into account).

to continue in Year 2, the performance metric used (revenue or tax assessment) was assigned to be the same in Year 2 as it was in Year 1. Note also that an additional 115 new circles were included for the first time in the Year 2 lottery.

Lotteries were conducted by computer publicly in the central tax authority office in Lahore at the start of each fiscal year.[19] Appendix Table A.1 compares treatment circles to control circles on key tax recovery variables at baseline. The treatments appear completely balanced in Year 1 (p-value 0.674); in Year 2; the p-value for balance is 0.065; and pooled, the p-value is 0.230. We have verified that controlling for all these variables does not qualitatively affect the main results (see Appendix Table A.3.)

4.2 Estimating Specifications

To test whether the incentives embodied in the transfer mechanism outlined above actually led to improved performance, we estimate treatment effects on log revenue for circle c as follows

$$\log y_{ct} = \alpha_t + \gamma_t \log y_{c0} + \beta TREAT_c + \epsilon_{ct} \tag{6}$$

where $\log y_{c0}$ is the baseline value of the outcome variable and $TREAT_c$ is a dummy for being in the first year of receiving a treatment. In estimating equation (6) for Year 2, we restrict ourselves to circles that were randomly selected to be in the control group in Year 1, so β from equation (6) can be interpreted as the pure incentive effects of the scheme, before any allocations have taken place. We estimate the equation separately for Year 1, Year 2, and pooling both years together (with time fixed effects α_t and separate coefficients γ_t for each year in the pooled regression).[20] We report randomization-inference based p-values, where we use our actual randomization Stata code (including both assigning circles to groups, and then assigning groups to treatment or control) to generate counterfactual randomizations. We explore time dynamics and allocation effects below.

The theoretical analysis in Section 3 has predictions for which inspectors would face the highest marginal returns under incentive scheme, as given by equation (4). We use the baseline preferences elicited from inspectors, and the simulations described in Section (3.2), to compute the predicted equilibrium effort $\tilde{e}_i$ for each inspector i under the model.

We then test whether those inspectors predicted to have higher marginal incentives under PRSD do in fact respond more when randomly allocated to the treatment by estimating the following

[19] In Year 1, the lottery for Lahore to assign circles to groups was held on July 26, 2013; baseline preference data was collected between July 27 and July 31, and the lottery to assign groups to treatment or control status was held on August 3. Outside of Lahore, the lottery to assign circles to groups was held on August 3, 2013; baseline preference data was collected between August 4 and August 20, and the lottery to assign groups to treatment or control status was held on August 29. In Year 2 the lottery to assign to assign groups to treatment was conducted province-wide on August 5, 2014.

[20] For our main results, we focus the 410 circles that participated in the Year 1 lottery, as this holds the sample constant throughout and as we only have baseline preference data for these 410 circles. Results for Year 2, however, are similar when we incorporate the new circles that were not included in Year 1 lottery. See Appendix Table A.4.

equation:

$$\begin{aligned} \log y_{ct} &= \alpha_t + \alpha_g + \gamma_t \log y_{c0} + \\ &\quad + \beta_1 TREAT_c \times \tilde{e}_i + \beta_2 \tilde{e}_i + \beta_3 TREAT_c \times X_c + \beta_4 X_c + \epsilon_{ct} \end{aligned} \tag{7}$$

where $\tilde{e}_i$ is the model's prediction for how much effort inspector i would exert, as estimated in the previous section. We normalize $\tilde{e}_i$ from the model to have standard deviation 1. Note that $\tilde{e}_i$ is predicted for all inspectors, both in treatment and control, and differences in $\tilde{e}_i$ across inspectors i arise only from differences in their baseline preferences $\mathbf{P}$ and in y_0. We use year 1 data to fit the model and generate the $\tilde{e}_i$, and then estimate equation (17) using year 2 data so as to provide a more demanding "out-of-sample" test; we also show that using $\frac{dE[u]}{de_i}$, which comes from simulations using only baseline data on $\boldsymbol{P}$ and y_0, instead of $\tilde{e}_i$ produces very similar results. We include group fixed effects (α_g); since randomization occurred by group, this subsumes the main effect of treatment ($TREAT_c$). We also control for the interaction of $TREAT$ with other circle characteristics X_c that may be correlated with the cost of effort in that circle, in particular the size of the tax base at baseline and the baseline recovery rate (i.e. the log share of taxes assessed that are actually paid.), as well as the main effects of those characteristics X_c. The coefficient of interest is β_1, which captures whether the performance-ranked serial dictatorship treatment was more effective for those inspectors predicted by the theory to face stronger incentives under the scheme.

In addition, we present further analysis below that takes advantage of the Year 2 re-randomization to examine both the dynamic implications from running the PRSD scheme (both when it is applied only once and when it is repeated) as well as trying to separate out the various components that may be at play when the scheme continues over time. To do so, we estimate the following regression:

$$\begin{aligned} \log y_{c2} &= \alpha + \gamma \log y_{c0} + \beta_1 TREAT\_Y1_c \\ &\quad + \beta_2 TREAT\_Y2_c + \beta_3 TREAT\_Y1_c \times TREAT\_Y2_c + \epsilon_{ct} \end{aligned} \tag{8}$$

where $TREAT\_Y1_c$ is a dummy for having received the treatment in the first year, $TREAT\_Y2_c$ is a dummy for receiving the treatment in the second year, and $TREAT\_Y1_c \times TREAT\_Y2_c$ is a dummy for receiving a treatment in both years. The coefficient β_1 is thus interpretable as the effect on Year 2 collections of receiving treatment in Year 1 and NOT receiving it in Year 2, which captures both persistence of the scheme over time and whatever implications there may from having allocated inspectors according to the PRSD; the coefficient β_2 is the effect of receiving the treatment for the first time in Year 2 relative to pure controls, and the effect $\beta_1 + \beta_2 + \beta_3$ is the effect in Year 2 of receiving a treatment in both years relative to pure controls.[21]

[21] Note while that the coefficient on β_2 should be very similar to the coefficient in (6) in Year 2 data, since both estimate the effect of starting to receive the scheme in Year 2 relative to pure controls, they need not be mechanically identical since the estimated coefficient on baseline recovery, γ, will be slightly different between the two regressions.

5 Results

5.1 Effect of the first year of treatment

We first examine the impact of the PRSD scheme had any positive impact on overall tax collections. While in general one would expect the scheme to at least weakly increase effort (and hence tax collections), it is worth noting that there are other possible effects. For example, the scheme may shorten an inspector's time horizon in a circle, since some inspectors (particularly low-performing inspectors in popular circles) may now expect to be replaced. With shorter time horizons inspectors may choose to invest less in a given circle than they may otherwise.

The empirical results from estimating equation (6) are presented in Table 3, separately for current year tax revenue, arrears revenue (i.e. collections against past-due amounts from previous years), and total tax revenue.[22] In the first year (columns 1-3), circles in which inspectors that were told they would be reallocated at the end of the year based on their performance grew by about 4.9 log points higher than the control group. Compared with the control group's average growth rate of 11.7 percent, this represents a 41 percent higher growth rate than controls. For inspectors who were first included in the scheme in the second year the impact is even greater – 9.2 log points higher revenue, or about 30 percent higher growth rate than controls. The results are presented graphically, both as CDFs and PDFs, in Figure 4, and show that in Year 2 the effects are even more pronounced at higher quantiles.

We should emphasize that both the Year 1 and Year 2 effects are the incentive impacts of being in the scheme for one year on (different) randomly-selected groups of inspectors. Therefore the difference between Year 1 and Year 2 inspectors is not on account of the former having been exposed to the scheme for longer (we will examine longer term effects in subsequent sections), but rather reflect perhaps a different (better?) understanding and perhaps increased credibility of the scheme for inspectors who were included in the scheme in the second year.[23]

It is also worth emphasizing that these are purely incentive effects based on expected *future* postings – these are the effects on revenue in year t from being told at the start of year t that one's posting in year $t+1$ will be based on performance in year t. While the results are therefore based on inspectors' beliefs about the future allocations under scheme, we verify in Appendix A.2 that indeed the PRSD scheme did in fact increase the link between performance and the likelihood of receiving one's most preferred allocations (both in terms of higher performing inspectors moving to a more desired location or, if the inspector had a status quo preference, being less likely to have to move).

The magnitudes of the incentive effects from the PRSD are substantial. By way of comparison, the financial incentive schemes we studied in the same property tax context in Khan, Khwaja

[22]Note that the results on total collections in columns (1), (4), and (7) are not always an average of current and arrears collections as the baseline control variable differs in each column.

[23]The control group also grew at a faster rate in Year 2 due to a province-wide revaluation effort, conducted in both treatment and control areas, so it is possible the larger effect on revenues in Year 2 just reflects the fact that the scheme leads to an approximately similar percent increase in the growth of tax revenues in both years.

and Olken (2016), in which inspectors, constables, and clerks were together paid an average of 30 cents for each marginal dollar of revenue they collected, increased total revenue collected by 9.4 log points in the second year they were in effect, and the most effective of the three incentive schemes we studied – a piece rate scheme – increased revenue by 12.9 log points. The performance-ranked serial dictatorship, studied here, increased total tax revenue by 9.2 log points – about three-quarters as large an effect as the maximally effective financial reward scheme we studied, which paid purely based on revenue collected. While the performance-ranked serial dictatorship may entail administrative and political costs, financially it was completely free to the government, whereas the financial incentives had the government almost doubling the wages of tax staff.[24] This suggests that leveraging postings for incentive purposes can be an extremely cost-effective way for the government to improve performance.[25]

Appendix Table A.2 disaggregates the results on overall tax revenue into components of tax revenue (the tax base, the net exemption rate, and the recovery rate)[26], as well as separating the results by whether the rank-ordering was done based on growth of tax revenue, or whether it was done based on the growth in tax base. Overall, the main channel through which inspectors increase tax revenue is by increasing the tax base, consistent with our findings in Khan, Khwaja and Olken (2016). The point estimates suggest somewhat stronger effects both overall and on the tax base for the treatment that directly incentivized revenue collection, which was perhaps most easily understood for inspectors, though the differences are not statistically significant (p-value 0.16 for revenue; 0.215 for tax-base). We pool both sub-treatments for the remainder of the analysis.

5.2 Heterogeneity by marginal return to effort

While the scheme on average leads to positive incentives for inspectors to exert effort and in turn to higher tax collections, recall that the theoretical analysis in section 3 implied that not all inspectors may face equally strong incentives given heterogeneity in baseline preferences and heterogeneity across circles. We now directly test for this heterogeneity by examining whether those inspectors predicted to face higher incentives under the scheme as predicted according to the model (see computation in Section 3.2) do in fact respond more when randomized into the treatment group.

We use Year 1 to estimate the model, and then use Year 2 as an out-of-sample check to see the extent to which the heterogeneity predicted by the model indeed matches heterogeneous responses to the experiment. The results, calculated by estimating equation (7) for Year 2 data, are presented in Table 4. Recall one important factor in this exercise is what one assumes about the knowledge

[24]While one might argue that the government might have to compensate the inspectors for the extra effort they exert under the scheme, anecdotal evidence suggest that inspectors are already earning substantial rents from the job, so this is unlikely to be the case in practice.

[25]Note that in the case of the performance-ranked serial dictatorship reducing tax evasion, the change in revenue for the government we estimate is actually the true increase in social welfare if we assume that the utility cost from the effort exerted by tax inspectors is small relative to the change in tax revenue. See Feldstein (1999), Chetty (2009), and the related discussion in Khan, Khwaja and Olken (2016).

[26]The net exemption rate is the ratio of tax bills after exemption to tax base, and the recovery rate is the ratio of actual revenue to the tax bills after exemptions. All variables are expressed in logs, so that we can decompose log revenue as follows: $\log revenue = \log taxbase + \log netexemptionrate + \log recoveryrate$.

inspectors have of each others' preferences as well as "business as usual" performance in the tax circles. We therefore present results in Panels A-D under different plausible assumptions regarding what inspectors know.

Panel A first presents the results where $\tilde{e}_i$ – the Nash equilibrium level of effort under the model – is calculated assuming inspectors know both the full vector of preferences $\mathbf{P}$ and predicted $\mathbf{y}$ for all inspectors in their group; Panel B is calculated assuming they know $\mathbf{y}$ but they assume that other inspectors' preferences $\mathbf{P}$ are random; Panel C is calculated assuming they know $\mathbf{y}$ but they assume that everyone has the same $\mathbf{P}$ that they do; and Panel D is calculated assuming they know P but not $\mathbf{y}$. We re-normalize $\tilde{e}_i$ to have standard deviation 1 in each panel so the magnitudes are comparable across panels. As one might expect circle-specific factors such as tax base to be correlated with $\tilde{e}_i$ and to affect the heterogeneity in terms of how costly effort could be, column 2 incorporates controls X_c and their interactions treatment $TREAT_c \times X_c$, where X_c is circle-level log gross demand (i.e. the total tax base in the circle) and circle-level recovery rate (i.e. what share of taxes are recovered) at baseline (given that team size remains the same regardless of circle attributes, larger circles and circles with already higher recovery rates will likely require more effort to generate further increases in revenue).

The results show that inspectors do indeed respond to the transfer-based incentive scheme more when predicted to do so. Panel A suggest that one standard deviation higher predicted effort from the model increases the average treatment effect on current-year tax collection by 0.038 log points if no cost variables are included, and by 0.042 log points when the cost variables are included, but these results are not statistically significant.

Comparing across the panels in Table 4 we can see that the inspectors' beliefs that are most consistent with our results is the scenario in Panel B where inspectors have full knowledge of predicted differences in outcomes $\mathbf{y}$ (i.e. they do as good a job as predicting business-as-usual performance as we do using data from past years growth rates) but know little about others' preferences.[27] In particular, the results in Panel B are that one standard deviation higher predicted effort from the model increases the average treatment effect on current-year tax collection by 0.076 log points if no cost variables are included (p=0.033), and by 0.176 log points when the cost variables are included (p=0.08). The fact that inspectors seem to know something about predicted y is not surprising, since the main predictor of $\mathbf{y}$ is the previous year's tax recovery level, which is public information (see Appendix Table A.8). The results assuming full knowledge of predicted differences in outcomes $\mathbf{y}$ but assuming random preferences are at least twice as strong as the results that come from assuming full knowledge of both $\mathbf{y}$ and $\mathbf{P}$. Assuming everyone has the same preferences appears to be a poor assumption and is quite strongly rejected in the data.

Note that these are not mechanical effects, not only because the model is fit using year 1 data and predicted on year 2 data, but also in the sense that the only information from the treatment used in calculating the $\tilde{e}_i$ vectors is a single moment – the average impact of the treatment as

[27] We should note that we are not making claims about overall model fit – in fact the R^2 is very similar and high across all models (perhaps unsurprising given we include baseline value of the dependent variable in these regressions). Rather, we are highlighting the model where we find relatively robust and significant interaction effects.

estimated in Table 3; no information about heterogeneous responses is used to calculate $\tilde{e}_i$.[28]

It is also instructive to conduct an exercise where we estimate heterogeneity of treatment effects based only on *ex ante* data (such as the distribution of $\mathbf{P}$ and $\mathbf{y_0}$, and circle characteristics). This exercise is important because, to the extent that it works, it implies that the model in Section 3 can be used to predict whether a particular service or unit would be a good candidate for application of a performance-ranked serial dictatorship *before* applying the scheme. To do so, we re-estimate equation (7), but instead of using the model-based equilibrium effort vectors $\tilde{e}$, we instead use the marginal incentives $\frac{dE[u]}{de_i}$ across inspectors i, calculated under the model evaluated at the point $\boldsymbol{e} = \mathbf{0}$, as discussed in Section 3.2 above (again standardized to have mean 0 and standard deviation 1). Importantly, this measure of marginal returns to inspector effort is calculated entirely using *ex ante* data. The results are presented in columns 3 and 4 of Table 4, and are virtually identical to the results in columns 1 and 2 estimated using the full equilibrium effort vector $\tilde{e}$..

In sum, the results suggest that inspectors seem to have some reasonable understanding of their marginal incentives induced by the scheme, and to respond accordingly. These results also provide an empirical validation for the theory outlined above: the theory appears to have some predictive power as to which inspectors respond to the incentives most. And, the results imply that if one knows the distribution of $\mathbf{P}$ and $\mathbf{y}$, the model in Section 3 can be used to predict whether a particular service or unit would be a good candidate for application of a performance-ranked serial dictatorship.

5.3 Preferences and allocation effects

The results thus far have focused on the incentive effects of the scheme. But, the scheme also changes allocations of inspectors to circles since best-performing inspectors are given priority in choosing where they should be located. We have already noted previously (and detailed in Appendix A.2) that the PRSD scheme did in fact increase the likelihood that higher performing inspectors will move to/stay in their desired locations. To understand the allocation changes induced by the PRSD scheme more generally, this section explores the preferences of inspectors and how the scheme affects the types of circles to which high-performing inspectors are allocated.

We begin by exploring what attributes tend to make a circle popular – both for typical inspectors, and for the inspectors who end up being ranked highly (and who therefore are more likely to receive their top choices). Column 1 of Table 5 examines how top-ranked circles in Year 1 (based on preferences expressed by all inspectors at baseline) compare to the average circle, on a variety of metrics also measured at baseline. We draw circle characteristics both from our administrative data, and from a property survey of over 16,000 properties we conducted at the time this experiment

[28]Note that the $\tilde{e}_i$ vectors on the right-hand side of equation 7 depend on the estimate of α, which could in principle introduce additional uncertainty beyond that captured by the standard errors shown in Table 4. In practice, however, since the standard error on the estimate of α is so small (see Appendix Table A.9), the additional variance introduced by the uncertainty around α is virtually non-existent. To verify this, we reproduce Table 4 using estimates of $\tilde{e}_i$ derived at the 95 percent confidence interval of α (that is, $\hat{\alpha} + 2se$ and $\hat{\alpha} - 2se$) in Appendix Table A.7, and show they are virtually identical to the estimates in Table 4.

was starting.[29] Column 2 goes even further, and compares how circles ranked first by the ranked inspector in each group – i.e. the inspector who will get his first choice no matter the preferences of others – compare to the typical circle. Each cell reports a separate univariate regression comparing top-ranked circles to average circles.

The results show that inspectors appear to prefer circles that are large (i.e. Column 2 shows that the top ranked inspectors are choosing circles that have a tax base 34 log points larger than average). Their more preferred circles have fewer properties, but more valuable ones – perhaps making for an easier job. These properties actually have lower bribe rates (where the bribe rates are calculated as log of the the ratio between the typical bribe given (measured from the survey) and the average property value).

The remaining columns repeat the same exercise, but for the allocations induced by the scheme so we can see what (differential) allocation occurred in equilibrium as a result of applying the scheme. Column 3 compares the circles where the top ranked inspector in every treatment group ended up to the typical circle. Specifically, we define y_{top} as the value of y for circles where top-ranked inspectors are allocated, and then compute $E(y_{top} - \bar{y} \mid TREAT = 1)$. As expected, since the top-ranked inspector gets his first choice for sure – this looks very similar to column (2).[30]

Columns 4 and 5 show the treatment effect on allocations - i.e. how the treatment affected where top inspectors were placed (relative to control/business as usual). Specifically, for each characteristic y, we compute the difference in allocations of top inspectors between treatment and control, i.e.

$$E(y_{top} - \bar{y} \mid TREAT = 1) - E(y_{top} - \bar{y} \mid TREAT = 0) \tag{9}$$

In column (4), we restrict attention to treatment circles where ranking was done based on actual tax revenue, and calculate y_{top} for controls using tax revenue to rank control circles; column (5) analogously calculates y_{top} restricting treatments to those ranked using tax base and using tax demand to rank controls. All data is from the first year. These equations estimate how the scheme changes allocations compared to the allocations of top performers in the status quo.

When we focus on allocation effects estimated using equation (9) – two main results stand out. Column (4) shows that, for the circles ranked based on revenue, the scheme results in top-ranked inspectors being more likely to be allocated to larger circles. Column (5) shows that, for the circles ranked based on tax base, the scheme results in an allocation of inspectors towards circles with fewer, but higher value, properties. The results are largely similar if we consider results for the top-three ranked inspectors instead of the just the top-ranked inspector (see Appendix Table A.10).

The fact that there are some differences in allocations induced by the scheme may be important

[29] This survey was also the endline survey for Khan, Khwaja and Olken (2016); more details about the survey can be found there.

[30] Columns (2) and (3) need not be identical for two reasons. First, column (2) considers both treatment and control, whereas column (3) considers just treatment circles. While baseline preferences should be similar between treatment and control due to randomization, they will not be numerically identical. Second, inspectors were given the chance to confirm (and revise if necessary) their preferences in the third quarter of each year. The preferences used for assignment (and hence analyzed in column 3) end up being very similar, but not necessarily identical, to preferences expressed at baseline analyzed in column 2.

in the longer term. On the one hand, if one had the view that the tax department had previously been allocating people optimally to maximize the match between tax inspector characteristics and circle needs, then any deviation from the status quo might reduce the department's welfare. On the other hand, the scheme appears to induce an allocation that puts top-performing inspectors in larger circles and those with more valuable properties. If these individuals continue to perform better this reallocation would increase overall tax collection.

5.4 Dynamic effects from repeated application of the transfer scheme

The results thus far have focused on the first year the incentive scheme was in place. While we find relatively large impacts especially for those inspectors predicted to face higher marginal incentives, an interesting question to ask is whether and how these results could change were the scheme in place every year. One could imagine several reasons why these results could differ over time. First, it could simply be that there is a limited scope for improvement and the impact of the incentives diminishes over time. Second, the allocation effects discussed above could affect performance, both in terms of level (i.e. moving better people to bigger circles, as discussed above, could affect collections), and in terms of responsiveness to treatment (i.e., people may respond more to incentives in a place they have selected vs. one that was exogenously assigned). Third, there could be (adverse) disruption effects as people may perform differently when they have moved to a new place. Finally, there could also be differential investment effects – if inspectors think they are likely to be moved again quickly, they may not invest much in their new locations. And, knowing that they are only in a new position for a year, they may change their preference ratings – if, for example, there is a fixed effort cost of adjusting to a new location, those who know they may move again after a year may prefer to just stay in place rather than move again and again.

To examine these issues empirically, we re-randomized the scheme at the beginning of Year 2 (as described in Section 4). This created four groups, as shown in Table 2. To analyze the differential effects, we restrict ourselves to Year 2 data, and estimate equation (8), which separately estimates impacts for having received the PRSD scheme in Year 1 and not in Year 2 (β_1), for receiving it for the first time in Year 2 (β_2), and for receiving it in both Year 1 and Year 2 (given by $\beta_1 + \beta_2 + \beta_3$).

The results are presented in Table 6 for total recovery, current year recovery, and arrears. There are several interesting results here. First, for both total and current recovery, we cannot reject the null that $\beta_1 = \beta_2$, i.e. that the initial effect of having the program persists in the second year.[31] We should note that the difference between β_1 and β_2 captures not only differences in (i) persistence (i.e. the former is the impact a year after the scheme has ended, and the latter is the impact due to the incentives created by the scheme for the first time), but also (ii) any changes in understanding/credibility about the scheme (those newly entering the scheme in the second year have more information about how the schemes works, whether it is credible, and perhaps others

[31] Though the point estimates are quite similar, 0.109 for β_1 and 0.081 for β_2, the 95 percent confidence interval of $\beta_1 - \beta_2$ is substantial, and runs from -0.077 to $+0.133$. We can, however, reject the null that effects were zero for both β_1 and β_2.

preferences), and (iii) any differences between circle allocation of the type discussed in Section 5.3 (since those who first entered the scheme are likely to be in different circles now as compared to those who entered in the second year, since they have been re-assigned based on performance).

Nevertheless, on net, the key result is that the effects persist strongly even after the incentives have been turned off. Indeed, the point estimate of β_1, 0.109, is substantially larger than the Year 1 effect estimated in 6. This suggests that if anything, the allocation effects induced by the scheme may have further enhanced its effects, rather than diminished them.

Second, a key result is the negative interaction term on β_3. Note that we cannot reject that the net treatment effect of experiencing the scheme the the second time in a row is the same as experiencing it once (i.e. we cannot reject that $\beta_1 + \beta_2 + \beta_3 = \beta_2$, i.e. that $\beta_1 + \beta_3 = 0$ (p-value 0.401); what is clear though is that the effect of receiving the scheme twice is definitely not twice the effect of receiving it once.

There are several factors that may be at play here. First, the simulations of the model suggest that as inspectors learn each other's preferences, this will tend to dampen incentives. Although we do not have direct data on what inspectors believe about others, it is certainly plausible that they know more about other inspectors' preferences P in the second year, after observing a full set of allocations, than when in the scheme for the first time.

Second, recall that as part of the design inspectors were allowed to change their preferences before the first round of allocations occurred (but after they had found out whether they been re-selected for continuation in the scheme or not). Appendix Table A.12 shows that inspectors who know they will participate in the scheme again rate their own circle higher; that is, they are 14 percentage points more likely to prefer the status quo and not move positions if they know they will face the PRSD scheme again in Year 2. This suggests that one possible reason for the smaller effect is the allocations may differ. Another important difference between those inspectors randomly selected to receive the scheme twice is and those receiving it for the first time in Year 1 that those who experienced the treatment already in Year 1 may have been more likely to have been moved already compared to those who had not. This would be the case if the performance-ranked scheme creates more movements than occur as part of the status qu, and indeed, this apears to be the case – Year 1 circles were about 10 percentage points more likely to have experienced a move. [32] Being newly placed in a circle may make it harder to exert effort in response to the Year 2 treatment; a new inspector may not know, for example, which properties can be added to the tax rolls.[33]

A final explanation for the negative interaction effect is simply discouragement: inspectors may

[32] Appendix Table A.13 investigates this by looking both at a dummy for whether the inspector present the circle at the mid-point of Year 2 (i.e. after postings from the scheme had been implemented) was the same as the inspector who was in the circle at baseline, and also at the number of days that same inspector had been posted in the circle. The results show that the Year 1 circles were about 10 percentage points more likely to have experienced a move.

[33] The increased disruptions may also have had a direct negative effect on revenues, but this seems small. We explore this in more detail in appendix A.3, where we use baseline preferences and heterogeneity across circles in how "business as usual" revenue growth interacted with treatment as an instrument for being moved. Overall, the estimates suggest a negative effect of movements on total revenue of about 6 percent (statistically significant using OLS, but noisy using IV). While the noise in the IV estimates suggest interpreting them with caution, they indeed suggest that movements per se do adversely impact performance.

have felt that they just worked hard in Year 1 under the scheme, only to see their hard work be "for nought" in the sense that both the new posting was only for one year, and they need to work hard again in the second year. In any case, the results presented here suggest that while this type of incentive scheme can be effective, it cannot be applied year after year.

6 Conclusion

Effective state bureaucracies play a central role in facilitating growth and development (Bertrand et al., 2016; Best, Hjort and Szakonyi, 2017; Xu, 2017). Recent work in economics has explored the importance and challenges of providing financial incentives to bureaucrats, both at the selection stage and once on the job (e.g., Dal Bó, Finan and Rossi, 2013; Deserranno, 2015; Khan, Khwaja and Olken, 2016; Ashraf et al., 2016; Fisman and Wang, 2017). However, in many contexts, governments face many constraints on their ability to provide financial incentives, as many governments have adopted strict civil service rules in an attempt to limit politicians' ability to use government jobs to reward political cronies. In these systems, pay and promotion are often rigid and mechanical, usually based on initial level and seniority rather on performance, and introducing explicit financial incentives is difficult.

This paper explores an alternative yet feasible approach to providing incentives. Many governments informally use postings, or horizontal movements, as a feasible avenue for rewarding good performers and punishing bad performers. In practice, however, the ambiguity of assignment rules and issues with revelation of agents' preferences over postings may limit the degree to which these can provide *ex ante* incentives to improve performance.

We propose a strategy-proof mechanism, the performance-ranked serial dictatorship, for using lateral transfers to provide incentives within groups. We then show, using a randomized experiment carried out over two years in a real tax bureaucracy in Punjab, Pakistan, that formalizing the relationship between performance and transfers indeed improves performance. By the second year of our study, those tax inspectors randomly allocated to the performance-ranked serial dictatorship had a 41 percent higher growth rate in tax revenues than control tax inspectors. This is almost the same magnitude of impacts as a performance-pay scheme we previously evaluated in the same context, but rather than having to double inspectors' pay, the zero-sum transfer mechanism was virtually free for the government.

Our paper, in combination with our previous work (Khan, Khwaja and Olken (2016)), raises interesting questions regarding how best to utilize pecuniary and non-pecuniary incentive systems over the longer term. Both papers show that the primary means through which incentivized tax inspectors increase collections is by expanding the tax base. Regardless of whether this is due to a reduction in collusion or greater effort to uncover true tax liabilities, it does suggest that such schemes may have lasting benefits. Therefore, to the extent that these schemes are costly to implement - pecuniary schemes incur incentive payments and posting based schemes may induce disruption costs - this suggests one may want to introduce such schemes every few years. Moreover,

to the extent that there is collusion between taxpayers and tax collectors, it may be desirable to not have inspectors be able to maintain their status quo location indefinitely. Since there is more agreement on rankings once status-quo preferences are not allowed, and since the PRSD mechanism provides stronger incentives when there is more agreement on preferences, it may be even be the case that the posting-based mechanisms are more effective in these cases.

The potential downside of performance-based posting is that the principal loses some flexibility over assignment. In our case, preferences were such that the best inspectors tended to want to be placed in the largest circles, which seems consistent with the principal's likely objectives, but this type of alignment may not necessarily be the case in other contexts. For example, a school system may want to assign top teachers to disadvantaged (and perhaps less popular) schools, or a tax administrator may want to mandate that no inspector can stay in the same location for more than few years. One could potentially address these issues by placing restrictions on the preference set, which we regard as an interesting direction for future work. More generally, our results demonstrate that bureaucracies have tremendous potential to improve performance at little financial cost by periodically using postings as an incentive, particularly when preferences over postings have a substantial common component.

References

Abdulkadirŏlu, Atila, and Tayfun Sönmez. 1998. "Random Serial Dictatorship and the Core from Random Endowments in House Allocation Problems." *Econometrica*, 66(3): 689–701.

Ashraf, Nava, Oriana Bandiera, Scott S Lee, et al. 2016. "Do-gooders and go-getters: career incentives, selection, and performance in public service delivery." *STICERD-Economic Organisation and Public Policy Discussion Papers Series*, 54.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2013. "Team incentives: Evidence from a firm level experiment." *Journal of the European Economic Association*, 11(5): 1079–1114.

Bertrand, Marianne, Robin Burgess, Arunish Chawla, and Guo Xu. 2016. "The costs of bureaucratic rigidity: Evidence from the Indian Administrative Service." *Unpublished working paper. University of Chicago.*

Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2017. "Individuals and Organizations as Sources of State Effectiveness, and Consequences for Policy." National Bureau of Economic Research.

Boudreau, Kevin J, Karim R Lakhani, and Michael Menietti. 2016. "Performance responses to competition across skill levels in rank-order tournaments: field evidence and implications for tournament design." *The RAND Journal of Economics*, 47(1): 140–165.

Bull, Clive, Andrew Schotter, and Keith Weigelt. 1987. "Tournaments and Piece Rates: An Experimental Study." *Journal of Political Economy*, 95(1): 1–33.

Camerer, Colin F, Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A cognitive hierarchy model of games." *The Quarterly Journal of Economics*, 119(3): 861–898.

Chetty, Raj. 2009. "Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance." *American Economic Journal: Economic Policy*, 1(2): 31–52.

Crawford, Vincent P, and Nagore Iriberri. 2007. "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica*, 75(6): 1721–1770.

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

Deserranno, Erika. 2015. "Financial Incentives as Signals: Experimental Evidence from the Recruitment of Health Promoters." Northwestern University.

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

Fisman, Raymond, and Yongxiang Wang. 2017. "The Distortionary Effects of Incentives in Government: Evidence from China's "Death Ceiling" Program." *American Economic Journal: Applied Economics*, 9(2): 202–18.

Gibbons, Robert, and Kevin Murphy. 1992. "Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence." *Journal of Political Economy*, 100(3): 468–505.

Glicksberg, Irving L. 1952. "A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points." *Proceedings of the American Mathematical Society*, 3(1): 170–174.

Holmström, Bengt. 1999. "Managerial incentive problems: A dynamic perspective." *The Review of Economic Studies*, 66(1): 169–182.

Iyer, Lakshmi, and Anandi Mani. 2012. "Traveling agents: political change and bureaucratic turnover in India." *Review of Economics and Statistics*, 94(3): 723–739.

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

Lazear, Edward P, and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *The Journal of Political Economy*, 841–864.

Prendergast, Canice. 1999. "The provision of incentives in firms." *Journal of economic literature*, 37(1): 7–63.

Shapley, Lloyd, and Herbert Scarf. 1974. "On cores and indivisibility." *Journal of mathematical economics*, 1(1): 23–37.

Sï¿œnmez, Tayfun. 2013. "Bidding for Army Career Specialties: Improving the ROTC Branching Mechanism." *Journal of Political Economy*, 121(1): 186–219.

Sunde, Uwe. 2009. "Heterogeneity and performance in tournaments: a test for incentive effects using professional tennis data." *Applied Economics*, 41(25): 3199–3208.

Svensson, Lars-Gunnar. 1999. "Strategy-proof allocation of indivisible goods." *Social Choice and Welfare*, 16(4): 557–567.

Xu, Guo. 2017. "The Costs of Patronage: Evidence from the British Empire." *Working Paper.*

Table 1: Summary statistics

	Mean	SD	Mean of within-group SD	N
Log Revenue (Total)	16.12	0.79	0.67	518.00
Log Revenue (Current)	16.00	0.80	0.69	518.00
Log Revenue (Arrears)	13.54	1.20	0.90	514.00
Log Tax Base (Total)	16.45	0.82	0.65	518.00
Log Tax Base (Current)	16.29	0.79	0.65	518.00
Log Tax Base (Arrears)	14.05	1.43	1.08	514.00
Log Recovery Rate (Total)	-0.08	0.11	0.10	518.00
Log Recovery Rate (Current)	-0.08	0.10	0.09	518.00
Log Recovery Rate (Arrears)	-0.13	0.22	0.16	514.00
Log Non-Exemption Rate (Total)	-0.25	0.22	0.17	518.00
Log Non-Exemption Rate (Current)	-0.22	0.17	0.13	518.00
Log Non-Exemption Rate (Arrears)	-0.38	0.58	0.45	514.00

Notes: Statistics from administrative data are shown at the end of Year 2 of the study (FY 2015). Each observation is one of the 525 circles as defined at the time of randomization.

Table 2: Treatment assignment of circles in Year 2

	Year 2 Control	Year 2 Treatment	Total
Year 1 Control	207	50	257
Year 1 Treatment	72	81	153
(Not included in Year 1 lottery)	96	19	115
Total	375	150	525

Table 3: Treatment Effect on Log Tax Revenue

	Year 1 (Y1 Q4)			Year 2 (Y2 Q4)			Pooled		
	(1) Total	(2) Current	(3) Arrears	(4) Total	(5) Current	(6) Arrears	(7) Total	(8) Current	(9) Arrears
Treatment	0.049	0.048	0.065	0.092	0.069	-0.074	0.061	0.054	0.026
	(0.022)	(0.023)	(0.056)	(0.042)	(0.040)	(0.119)	(0.020)	(0.021)	(0.052)
	[0.009]	[0.023]	[0.259]	[0.036]	[0.142]	[0.594]	[0.002]	[0.004]	[0.653]
Baseline	0.892	0.892	0.796	0.946	0.952	0.808	0.944	0.951	0.812
	(0.018)	(0.024)	(0.028)	(0.019)	(0.019)	(0.051)	(0.019)	(0.019)	(0.052)
N	405	405	396	251	251	244	656	656	640
Mean growth in controls	0.117	0.154	-0.048	0.309	0.408	-0.337	0.203	0.268	-0.177

Notes: Tax revenue (columns 1, 4, and 7) is comprised of revenue from the current years tax due (columns 2, 5, and 8), plus revenue collected from previous years' unpaid taxes (denoted 'arrears', columns 3, 6, and 9); the dependent variable in each column is the log of the respective tax measure. Estimation is by OLS. The unit of observation is a circle, as defined at the time of randomization. Specification controls for baseline values (FY 2013). Robust standard errors in parentheses. Standard errors are clustered by circle. Randomization inference based p-values in brackets.

Table 4: Heterogeneity in treatment effects in Year 2 by simulated marginal returns to effort

	Eq effort e		dEu/de at e=0	
	(1)	(2)	(3)	(4)
Panel A: Full knowledge of P, Y				
Treatment * Model-predicted effort	0.038	0.042	0.034	0.038
	(0.051)	(0.066)	(0.048)	(0.062)
Treatment * Tax base at baseline		0.004		0.005
		(0.061)		(0.061)
Treatment * Recovery rate at baseline		-0.078		-0.076
		(0.207)		(0.207)
Model-predicted effort	-0.017	-0.025	-0.016	-0.023
	(0.021)	(0.023)	(0.020)	(0.022)
R-squared	0.935	0.936	0.935	0.936
Panel B: Random P, full knowledge of Y				
Treatment * Model-predicted effort	0.076	0.176	0.076	0.178
	(0.036)	(0.103)	(0.031)	(0.087)
Treatment * Tax base at baseline		-0.081		-0.104
		(0.073)		(0.074)
Treatment * Recovery rate at baseline		-0.339		-0.377
		(0.310)		(0.292)
Model-predicted effort	-0.018	-0.046	-0.018	-0.044
	(0.038)	(0.045)	(0.033)	(0.040)
R-squared	0.936	0.937	0.936	0.938
Panel C: Assume identical P, full knowledge of Y				
Treatment * Model-predicted effort	0.025	0.051	0.027	0.055
	(0.027)	(0.051)	(0.028)	(0.051)
Treatment * Tax base at baseline		0.011		0.008
		(0.046)		(0.046)
Treatment * Recovery rate at baseline		-0.217		-0.230
		(0.297)		(0.301)
Model-predicted effort	0.006	-0.001	0.004	-0.004
	(0.027)	(0.028)	(0.028)	(0.029)
R-squared	0.936	0.936	0.936	0.937
Panel D: Full knowledge of P, no knowledge of Y				
Treatment * Model-predicted effort	-0.009	-0.014	-0.013	-0.019
	(0.060)	(0.069)	(0.056)	(0.065)
Treatment * Tax base at baseline		0.031		0.033
		(0.062)		(0.062)
Treatment * Recovery rate at baseline		-0.043		-0.041
		(0.181)		(0.180)
Model-predicted effort	-0.015	-0.018	-0.011	-0.015
	(0.019)	(0.020)	(0.018)	(0.019)
R-squared	0.935	0.936	0.935	0.936
N	249	249	249	249
Mean of control group	16.268	16.268	16.268	16.268

Notes: OLS regessions of log recovery on treatment assignment, with group fixed effects (Y2). The unit of observation is a circle, as defined at the time of randomization. In Columns 1 and 2, the model-predicted effort is the Nash equilibrium level of effort. In Columns 3 and 4, effort corresponds to the marginal incentives evaluated at e=0. Columns 2 and 4 include tax base and recovery rate at baseline and their interactions with treatment assignment in the specification. Robust standard errors in parentheses. Standard errors are clustered by circle.

Table 5: Preferences and allocations

	Y1 Preferences (Treatment)				Allocation		Difference in allocation			
	(1)		(2)		(3)		(4)		(5)	
	All circles		Top inspectors' circles		Treated inspectors		Treatment - Control (Revenue)		Treatment - Control (Tax base)	
	b / se	Mean	b / se	Mean	b / se	Mean	b / se	Mean	b / se	Mean
Log of tax base (Current)	0.167	15.870	0.343	15.873	0.312	15.906	0.537	16.055	0.173	16.050
	(0.070)		(0.177)		(0.179)		(0.304)		(0.278)	
Log of tax base (Arrears)	0.137	14.254	0.173	14.228	0.219	14.224	0.355	14.492	-0.092	14.552
	(0.128)		(0.413)		(0.407)		(0.677)		(0.667)	
Growth in tax base (Current)	0.001	0.101	0.004	0.099	0.011	0.094	0.006	0.113	-0.024	0.109
	(0.008)		(0.035)		(0.036)		(0.073)		(0.040)	
Growth in tax base (Arrears)	0.055	-0.321	-0.117	-0.335	-0.068	-0.361	-0.022	-0.362	-0.199	-0.317
	(0.086)		(0.199)		(0.220)		(0.414)		(0.216)	
Log of revenue (Current)	0.180	15.565	0.376	15.566	0.338	15.605	0.635	15.737	0.235	15.735
	(0.072)		(0.177)		(0.172)		(0.309)		(0.304)	
Log of revenue (Arrears)	0.151	13.848	0.113	13.814	0.152	13.821	0.669	14.023	-0.193	14.086
	(0.123)		(0.328)		(0.337)		(0.626)		(0.430)	
Growth in revenue (Current)	-0.003	0.142	0.024	0.140	0.029	0.138	0.057	0.172	0.040	0.158
	(0.011)		(0.036)		(0.037)		(0.067)		(0.060)	
Growth in revenue (Arrears)	0.068	-0.331	-0.192	-0.351	-0.164	-0.359	0.144	-0.353	-0.355	-0.312
	(0.093)		(0.220)		(0.242)		(0.435)		(0.231)	
Any unofficial payment	0.050	0.395	0.040	0.395	0.034	0.404	-0.039	0.387	0.196	0.375
	(0.026)		(0.081)		(0.079)		(0.134)		(0.139)	
Log of unofficial payment rate	-0.043	0.704	-0.219	0.728	-0.211	0.705	-0.378	0.692	-0.402	0.698
	(0.041)		(0.128)		(0.122)		(0.237)		(0.218)	
Log average p.c. expenditure	0.066	8.614	0.097	8.611	0.082	8.631	0.141	8.652	0.262	8.625
	(0.046)		(0.096)		(0.101)		(0.157)		(0.171)	
Properties for commercial use	-0.004	0.322	-0.072	0.325	-0.072	0.328	-0.016	0.367	-0.092	0.356
	(0.016)		(0.051)		(0.049)		(0.079)		(0.113)	
Properties for residential use	-0.006	0.424	0.114	0.419	0.119	0.413	0.056	0.377	0.150	0.381
	(0.015)		(0.068)		(0.068)		(0.102)		(0.153)	
Num of properties (in hundreds)	-5.497	65.585	-15.182	68.221	-11.301	63.547	-4.070	75.349	-30.497	74.123
	(3.594)		(7.654)		(6.780)		(14.674)		(11.300)	
Log of average property value	0.204	7.630	0.487	7.608	0.489	7.631	0.062	7.869	1.450	7.809
	(0.114)		(0.347)		(0.341)		(1.103)		(0.573)	
N	1184		136		123		197		199	

Notes: Columns 1 and 2 present OLS regressions of circles attributes on a dummy variable that takes the value of 1 for circles that were ranked as TOP 1. Sample consists in all treated circles and treated circles of TOP 1 inspectors, respectively. Column 3 shows regressions of circles characteristics on an indicator that takes the value of 1 if the treated inspector that ended up in that circle ranked among the TOP 1 of his group. Columns 4 and 5 report the difference in allocation between inspectors in the treatment and control group. Inspectors in Column 4 are ranked based on their performance in recovery (growth in recovery rate). In Column 5, based on their performance in demand (growth in tax base).

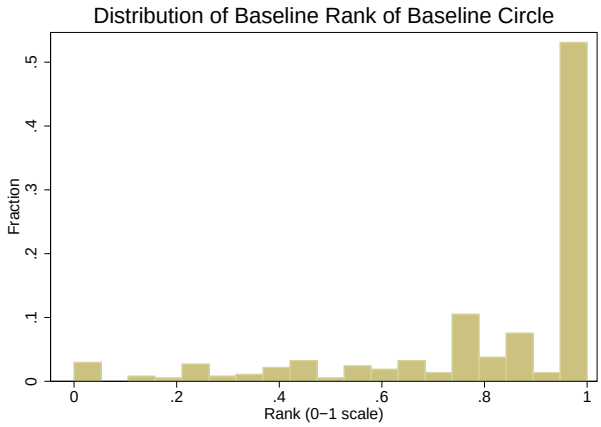
Table 6: Dynamic effects estimated in Year 2

	(1) Total	(2) Current	(3) Arrears
Y1 Treatment (β_1)	0.109	0.085	0.128
	(0.038)	(0.040)	(0.100)
	[0.003]	[0.020]	[0.200]
Y2 Treatment (β_2)	0.081	0.055	-0.074
	(0.043)	(0.041)	(0.119)
	[0.064]	[0.235]	[0.592]
Y1 AND Y2 Treatment (β_3)	-0.150	-0.085	-0.061
	(0.067)	(0.068)	(0.178)
	[0.014]	[0.203]	[0.733]
N	403	403	392
$\beta_1 = \beta_2$	0.564	0.560	0.167
$\beta_1 + \beta_3 = 0$	0.401	0.999	0.655
Mean growth in controls	0.309	0.408	-0.337

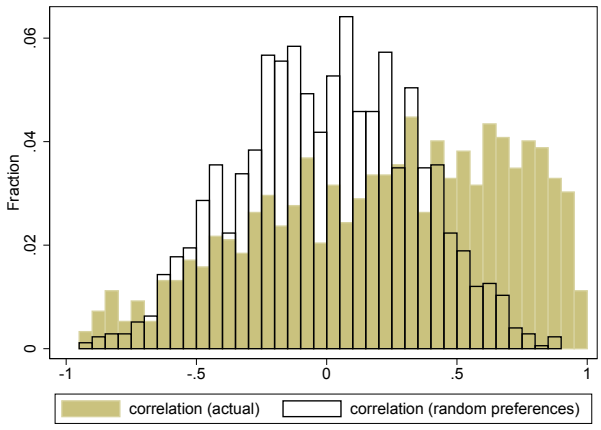
Notes: OLS regessions of log of tax revenue on Y1/Y2 treatment interactions The unit of observation is a circle, as defined at the time of randomization. Specification controls for baseline values. Robust standard errors in parentheses. Standard errors are clustered by circle. Randomization inference based p-values in brackets. RI p-values displayed for tests of equality of coefficients.

Figure 1: Descriptive statistics of baseline preferences over positions

(a) Distribution of inspector's rank of their status quo circle



(b) Distribution of pairwise rank correlations

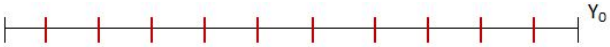


Notes: Figure 1a shows the histogram of inspectors ranks of their status quo circle, at baseline, where the top-ranked circle is normalized to 1 and the bottom ranked circle is normalized to 0. Figure 1b shows the distribution of pairwise rank correlations among inspectors within a given group. The histogram in outline shows what the distribution would look like if inspectors' preferences were random.

Figure 2: Different scenarios for the distribution of y_0 within a group

(a) When y_0 is concentrated, the marginal return to effort is high for all inspectors.

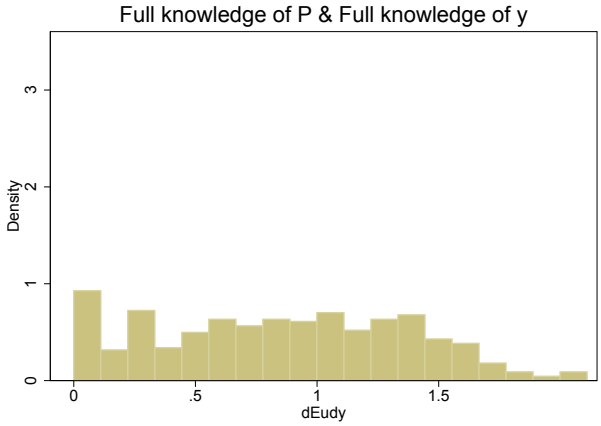
(b) When y_0 is spread out, marginal returns to effort are low.



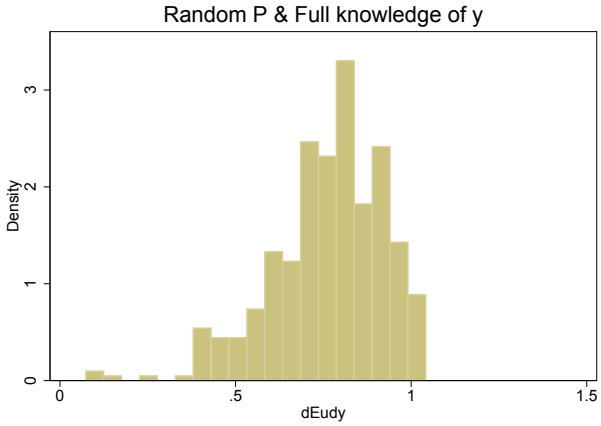
(c) Within a group variation: inspectors with y_0 close together face strong incentives, whereas those with y_0 far apart face weaker.

Figure 3: Simulated Distribution of $\frac{dE[u]}{de_i}$ under alternative assumptions about knowledge

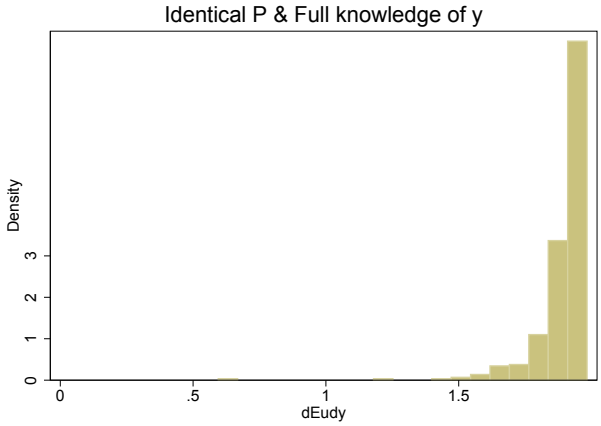
(a) Full Knowledge of P and y



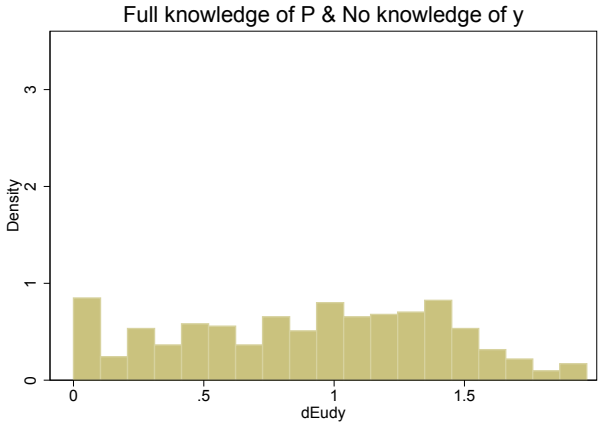
(b) Assuming random preferences P, full knowledge of y



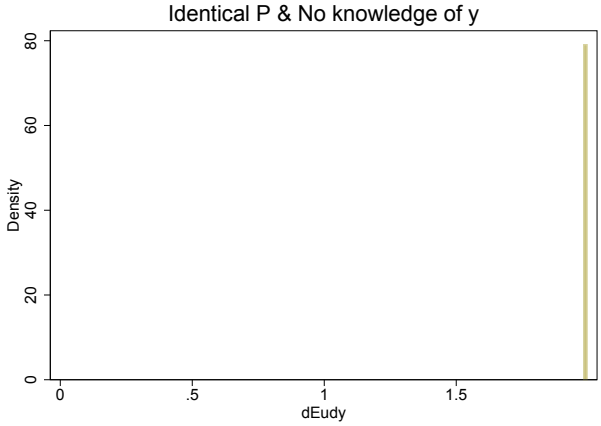
(c) Assuming identical preferences P, full knowledge of y



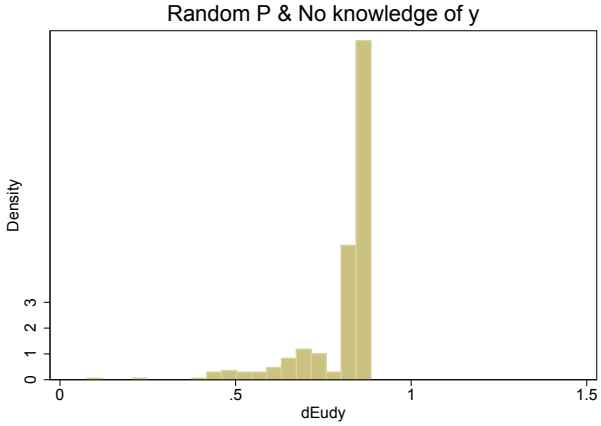
(d) Full knowledge of P, no knowledge of y



(e) Assuming identical preferences P, no knowledge of y



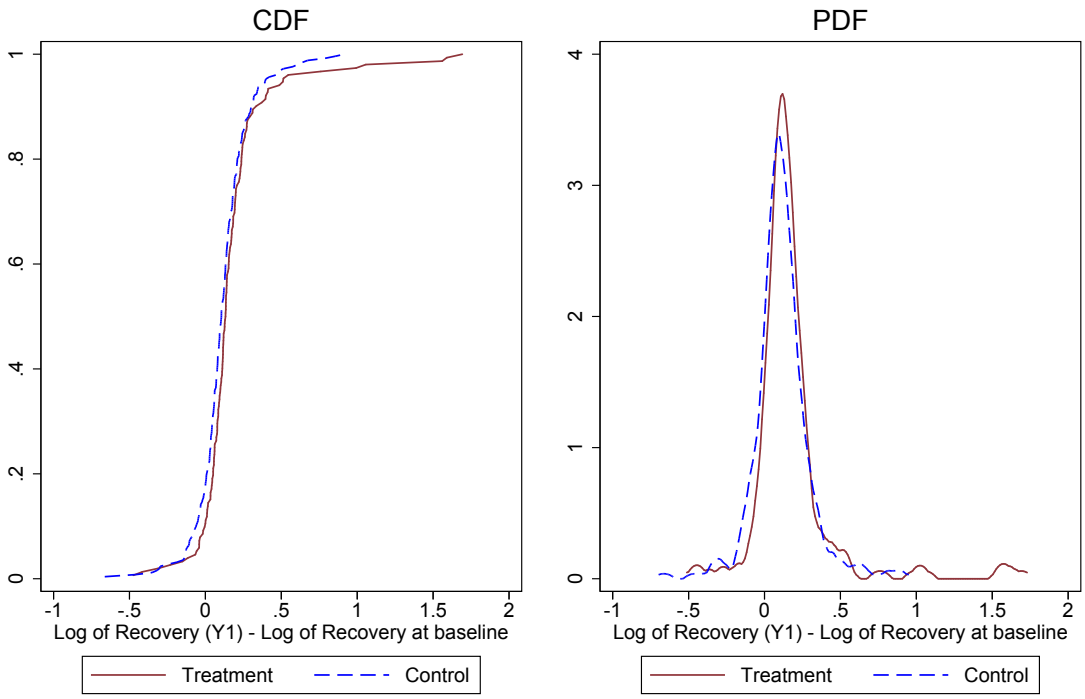
(f) Assuming random preferences P, no knowledge of y



Notes: Each figure shows the distribution of $\frac{dE[u]}{de_i}$ evaluated at $\boldsymbol{e} = 0$ under the knowledge assumptions stated. Simulations are as described in Section (3.2).

Figure 4: Distribution of change in log tax revenue collected, by treatment

(a) Year 1



(b) Year 2