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

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

July 8, 2016

Abstract

Transfers are often used by bureaucracies, especially in emerging economies, in an attempt to reward or punish their staff. Yet we know little about whether, and how, transfer mechanisms can help incentivize performance. Using transfers to induce performance is challenging, as heterogeneity in preferences over which postings are desirable non-trivially impacts the effectiveness of such schemes. We propose and examine the properties of a mechanism, which we term a *performance-ranked serial dictatorship*, in which individuals sequentially choose their desired location, with their rank in the sequence based ontheir performance. We then evaluate the effectiveness of this mechanism using a two-year field experiment with over 500 property tax inspectors in Punjab, Pakistan. We first show that the mechanism is effective: being randomized into the performance-ranked serial dictatorship leads inspectors to increase the growth rate of tax revenue by between 44 and 80 percent. We then use our model, combined with preferences collected at baseline from all tax inspectors, to characterize which inspectors face the highest marginal incentives under the scheme. We find empirically that these inspectors do in fact increase performance more under the scheme. We estimate the cost from disruption caused by transfers to be small, but show that applying the scheme too frequently can reduce performance. On net the results suggest that bureaucracies have tremendous potential to improve performance by periodically using transfers as an incentive, particularly when preferences over locations have a substantial common component.

*This project is the result of collaboration among many people. We thank Parag Pathak for helpful comments . We thank Gabriel Kreindler, Alyssa Lawther, Ismail Khan, Kunal Mangal, Wayne Sandholtz, Mahvish Shaukat, and Gabriel Tourek for outstanding research assistance in Cambridge and Zahir Ali, Osman Haq, Turab Hassan, Zahra Mansoor, Obeid Rahman, Ali Abbas, Shahrukh Raja, Adeel Shafqat, and Sadaqat Shah for outstanding research assistance in Lahore. We thank all the Secretaries, Director Generals, Directors, the two Project Directors from the Punjab Department of Excise and Taxation, the Punjab Finance, Planning and Development departments and the Chief Secretary and Chief Minister's offices for their support over the many years of this project. Financial support for the evaluation came from 3ie, the IGC, and the NSF (under grant SES-1124134). 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 highly regulated by 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 for career concerns (a la Gibbons and Murphy 1992; Holmström 1999) to provide incentives as well.

In such such rigid 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 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 transfers 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 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 to the extent that 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 person i may be a terrible posting for person j. To use postings as an incentive, managers need to take these diverse preferences into account. And in so doing, managers face 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 incentives. And even if the incentive aspects of such as 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.

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 transfers will be explicitly based on performance (and where they know this in advance), or into a control group where transfers operate as in the status quo (which we show empirically has little to do with objective performance measures).

In order to do so, we designed and implemented a novel, strategy-proof transfer mechanism, which we term a *performance-ranked serial dictatorship* (henceforth, PRSD). We build on the theoretical literature 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 for efficiently allocating a fixed set of slots to individuals when individuals' preferences are unknown. In general, however, this literature is silent 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 incentives to increase effort embodied in such a mechanism are complex and heterogeneous across individuals. In particular, the incentive effects depend on an individuals' own preferences among slots, the preferences of others, and the expected distribution of performance outcomes. 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.

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 model using rich data we collected 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 500 tax units, or 'circles.' Each circle covers a pre-defined geographic area, and is staffed by an 'inspector'. Within each district (i.e. roughly the same metropolitan area), we randomly assigned circles into groups of about 10 circles each. There was substantial heterogeneity within groups in circle characterisics – 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 they are clearly more "popular" circles that many inspectors like, there is also substantial idiosyncracies in preferences, including a substantial status quo bias, 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. (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; more on this below.) At the end of the year, postings within the group were re-assigned based on preferences, as per the mechanism. Control group transfers continued under business-as-usual. Groups were re-randomized at the beginning of year 2, 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 2nd year treatment groups at the end of year 2.

We find that overall, the promise of performance-based transfers substantially raised revenues. We estimate that revenues were about 5 log points higher in treatment groups than in control groups in first year, and 8 log points higher in treatment groups than in control groups in the second year. This amounts to an increase in the growth rate of tax revenues 44 percent in the first year and 80 percent in the second year. Note this is a pure incentive effect – this is the effect on revenue merely from announcing that the scheme will be applied to determine transfers in the subsequent year. These magnitudes are substantial compared to a performance pay scheme we evaluated in the same context: a previous randomized trial we conducted showed that paying the tax inspectors an average of 10 percent of every marginal dollar collected led increased tax revenues by about 9 log points (Khan, Khwaja, and Olken, 2016). Groups performed broadly similarly regardless of whether they were incentivized based on revenue or tax assessments.[1]

We then take the theory to the data to see whether or not those workers whom the theory predicts should face stronger incentives under the performance-based serial dictatorship do in fact respond more to the scheme. In particular, we use the model and simulate the marginal returns to effort implied by the performance-based serial dictatorship, 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 substantial evidence that those individuals who were predicted to have stronger incentives indeed increase their revenue more in response to being subject to the performance ranked serial dictatorship scheme. Of the two factors that predict marginal incentives, inspectors appear to be primarily using the second factor (forecasting their likely rank in the distribution), rather than information derived from the composition of preferences.These results imply that collecting information on 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 is also – to our knowledge – the first real world evidence outside of sports settings (e.g. golf) on how tournament-based compensation schemes can reduce effort for individuals who can forecast that they face small incentives from the tournament (Prendergast, 1999).

[1]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 quite similar in this context.

These effects thus far have focused only on the year the program was announced – when the program had incentive effects, but the allocation effects had not been realized. By re-randomzing treatments 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 transfers 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. The results are no greater, however, suggesting that allocating people to their preferred slots does not substantailly raise revenue, either.

We also one cautionary note, however: the effects disappear in the second year if an inspector is (randomly) exposed to the scheme for two years in a row. On net, these results suggest that a) it is predominantly the incentive effects that appear to matter, rather than the allocation efffects, as the effects appear immediately once incentives are put in place and don't increase after allocations have taken place, but b) these incentives can only be applied periodically, as applying them too frequently diminishes their effectiveness. ,

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

2 Setting, Data and Intervention

2.1 Property Tax Administration in Urban Punjab, Pakistan

Punjab is Pakistan's most populous province: its population of over 80 million would rank fifteenth in the world were it a country. 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). More details about the tax system can be found in Khan, Khwaja, and Olken (2016).

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

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 performance pay equal to a total of 30 percent of all taxes collected above a historical benchmark increases taxes by 9 log points. However, even though Khan, Khwaja, and Olken (2016) shows that it is cost-effective, given the substantial costs of performance pay to the government, after the experimental period ended they have yet to adopt it as a policy.

Instead, with limited reward mechanisms or vertical mobility, transfers – either to better or worse assignments – are the primary tool available to supervisors who want to improve performance. There is substantial heterogeneity in circles in terms of number of properties – for example, the 90th percentile circle has more than three times as many taxabe properties as the 10th percentile circle. Even more important is heterogeneity in ease of collecting taxs, opportunities for corruption, and amenities, all of which can be used to provide incentives. However, in practice, the transfer process is opqaue 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 transfers 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 datasource 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 and 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 verifications.

Summary statistics for key variables from the administrative data are shown in Appendix Table 9 for the second year of the experiment (FY 2015). First, current year revenues are substantially larger than arrears (i.e. collections against past years' unpaid taxes) – the mean of log current revenues is 16.00 compared with just 13.54 for log arrears, implying that, on average, current revenue is about 12 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 non-payment, a substantial issue is that many properties are either under-assessed or not assessed at all (see Khan, Khwaja, and Olken (2016) for detailed discussion of this issue.) Tax inspectors can therefore respond to performance incentives by adding new properties to the tax rolls, more accurately assessing existing properties, 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 10 circles from within the same metropolitan area. (More details on the construction of these groups below.) Inspectors 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 circle. [2]

Before beginning our analysis of incentive effects, it is useful to examine the preference data a bit further. We find that while there are clearly popular circles, inspectors also have idiosyncratic preferences. One way to see this is by examining inspectors' preferences for their status quo circle. 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.

Another way to characterize the ranks is to examine the distribution of average rank for circle j, averaging among over all inspectors i. This distribution if shown in Figure 1b. To interpret this, it is useful to consider two counterfactual distributions. First, Figure 1b also shows, in ouline, what this distribution would look like if inspectors preferences were *iid* random. Second, if inspectors completely agreed on the distribution of preferences for circles, the distribution would be uniform. The results in Figure 1b show that there is some agreement among inspectors – the distribution of average ranks is much more spread out than would be predicted simply by random chance. Combined, these two figures suggest two salient facts about the preference distribution: there is a substantial but not complete 'status quo bias', but there is also a non-trivial degree of agreement over which circles are most desirable. This common component of preferneces will help generate strong incentives under the PRSD scheme, as we discuss in more detail below.

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

[2]Inspectors had incentives to reveal their true preferences. The scheme was explained briefly to inspectors, so they could understand that truthful revalation 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).

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 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 up 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 heterogenous 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 & Taxation department for a two year period 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 transfers. 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 10 major metropolitan areas in Punjab, we randomly allocated circles into groups of approximately 10 circles each.[3] At the beginning of the tax year (i.e. in July), groups were randomly selected to either participate in the MBTP scheme or 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 told 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.[4] Performance was calculated 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' recovery 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.[5] Transfers were then carried out as described: the top-rank inspector

[3]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 areas for the purposes of forming groups.

[4]In order to convince inspectors in the first year that transfers would be made as promised, an additional group of 10 inspectors was randomly selected to have the merit-based transfers implemented at the start of year 1 based on performance in the previous year following the same performance-ranked serial dictatorship scheme. Transfers were indeed implemented as promised.

[5]Although final transfers were made on the basis of total performance during the fiscal year, inspectors were given

was given his first choice of posting, the second-rank inspector was given his top choice among remaining circles, and so on.

Note that transfers were done within groups, which as described above were randomly selected groups of 10 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 10 circles was for experimental feasibility, so that there would be both treatment and control areas within each metropolitan area. While 10 circles still entails substantial heterogeneity – within groups, the 90th pecentile circle has tax revenue almost nine times larger than the 10th percentile circle –the incentive effects we find here are most likely an under-estimate of the incentive effects that would be generated if choice was given over a larger number of locations.

3 Modelling Incentives under PRSD

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

3.1 Allocation under the Performance-Ranked Serial Dictatorship 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.

For any vector of outcomes $\mathbf{y}$, the performance-ranked serial dictatorship 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 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

information at the end of the third quarter as to their tentative ranking before submitting their final preferences.

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 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 is given by

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

The first-order condition suggests that there are several factors that 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.[6] 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. Without loss of generality, we can then re-label the circles j such that each inspector i's most preferred circle is $j = i$. In this case, the assignment function assigns inspector i to circle $j = i$ *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, 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.

[6] 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 1as 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.

A second factor that influences effort are the variances of the $\mathbf{y_0}$ and the error term ϵ_i in equation (1). To see this, consider the problem taking the realizations of $\mathbf{y_{-i}}$ as fixed, and continuing to assume that all inspectors have identical preferences $\mathbf{P}$. Order the realizations of $\mathbf{y_{-i}}$ from smallest to largest, and denote these as $z_1....z_{J-1}$, and let $z_0 = -\infty$ and $z_J = \infty$. Recall that $y_i = y_{i0} + e_i + \epsilon_i$, and assume that ϵ_i is distributed Normal with CDF Φ and PDF ϕ. Then $Pr(Rank(y) = j) = \Phi(z_j - y_{i0} - e_i) - \Phi(z_{j-1} - y_{i0} - e_i)$. Then the first-order condition in (5) can then be rewritten as

$$\sum_{\mathbf{j=1}}^{\mathbf{J}} u_{ij} \left[\phi(z_{j-1} - y_{i0} - e_i) - \phi(z_j - y_{i0} - e_i)\right] = c'(e_i) \tag{6}$$

To illustrate the impact of error variance, Figure 2 draws three examples of a equation (6) for a particular cutoff point, with low, medium, and high error variance. With very low error variance, the marginal return to effort is close to zero, since the inspector will end up ranked in between inspector with outcome z_{j1} and z_j almost for sure. For the medium-variance case, one can see a substantial return to effort, since a marginal increase in effort results in a substantial increase in the probability of landing ahead of the inspector with outcome z_{j-1} (given by $\phi(z_{j-1} - y_{i0} - e)$) and a relatively small increase in the chance of landing ahead of the next inspector (given by $\phi(z_j - y_{i0} - e)$). With very high variance, the marginal return to effort with respect to this particular cutoff point is also close to zero, because the PDF of the error distribution is almost identical at both cutoff points.

A third component that affects the returns to effort is information about $\mathbf{y_0}$. Heterogeneity in $\mathbf{y_0}$ – the predictable component of an inspector's performance – may affect incentives in complex ways, depending on how they interact with the utilities. Generally speaking, however, when the y_0 are very far apart (relative to the error variance), the same logic as before suggests that incentives will be weak. However, 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 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).

A final component that influences outcomes is the u_{ij}'s, the utility of different positions to the inspectors. In the simplest case with common preferences and common $\mathbf{y_0}$'s analyzed in equation (6), the classic result from Lazear and Rosen (1981) suggests that there exists a set of u_{ij}'s such that equation (6) 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 efficient. 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.[7]

Since the various components – information about alignment or idiosyncrasy of preferences $\mathbf{P}$, predictable performance y_0, different assumptions about 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 marginal incentives faced under the scheme given actual preference and (predicted) performance data in our context. We do so under different common knowledge assumptions that inspectors may hold about other's preferences and baseline performance. We do this in two steps. First, we approximate (4) assuming inspectors know both the baseline preference data submitted by all inspectors (in their group)and are able to predict $\mathbf{y_0}$ (i.e. growth rate that would be observed in the absence of effort) for all circles in their group. Second, we relax both of these knowledge assumptions in turn and simulate (4) under these different assumptions.

The details of the simulation exercise are presented in (A). Here we describe the basic outline of the process we use. In order to obtain $\mathbf{y_0}$, we first regress actual revenue change on two lags of a circle's (log) reveneue 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_ϵ^2 (the variance of the error in equation (1)). We then take 10,000 draws from the joint distribution of $\mathbf{y}$ given $\mathbf{y_0}$, as well as actual baseline preference data submitted by the inspectors, to calculate the empirical analogue of (4). We parameterize the utility function over difference circles (u_{ij}) linearlity, with $u_{ij} = 1$ for inspector i's top-ranked circle and $u_{ij} = 0$ for inspector i's lowest ranked circle.[8] Finally, to capture different levels of common knowledge, we consider the case of ignorance in predicted performance, by setting $\mathbf{y}_0 = 0$ for all inspectors, and/or the case of ignorance about preferences by assuming inspectors either believe that everyone else has the same preferences they do or that preferences are distributed randomly.

Panel A of Figure 3 plots a histogram of the distribution of $\frac{dE[u_i]}{de_i}$ across inspectors i ,assuming full knowledge of both $\mathbf{y_0}$ and $\mathbf{P}$. The figures shows substantial heterogeneity in marginal returns accross 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 plots the same figure under alternative assumptions of what inspectors know, turning off first knowledge of y (predicted performance), then knowledge of P, and

[7]While there are number of lab experiments along these lines (e.g. Bull, Schotter, and Weigelt 1987), we know of no field experiments in real-world settings that investigate the introduction of tournaments of any type, let alone comparing them to piece rate schemes.

[8]In the Appendix, we also consider two alternative utility functions, one that puts full weight on achieving ones top-choice and nothing otherwise (denoted $u2$), and one with full weight on achieving any circle strictly preferred to the status quo provides uility and zero othwewise (denoted $u3$)

then knowledge of both y and P. Note that when we turn off knowledge of P, we need to make an alternate assumption for what inspectors i believe about the preferences P for all other inspectors. We examine two possible assumptions: 1) that inspectors i believe that all other inspectors $-i$ have the same preferences they do, and 2) that inspectors i believe that inspectors $-i$ have random preferences. As discussed above and shown in Figure 1, preferences have both a common and idiosyncratic component, so reality is somewhere between these two extremes.

Note that less knowledge - either not knowing P or not knowing y – leads to a rightward shift in the distribution of expected utility. That is, adding knowledge about either P or y seems to dampen incentives for some people. Intuitively, with knowledge of P, some people know they are likely to get a good outcome regardless of how hard they work, dampening their incentives (panel b). With knowledge of y, 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 c and d).

Finally, if inspectors know neither P nor y, all inspectors in a given group have the same incentives. When inspectors assume that all inspectors share their preferenes, the incentives are 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 thier performance (e.g. if they uniquely prefer a given circle that everyone else ranks poorly). 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.

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 euqations 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

The primary empirical strategy used is to implement 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.[9]

In the second year, groups were re-randomized into treatment and control.[10] 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 will be continuing 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 1. Note that if a year 1 group was randomized 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.

Lotteries were conducted by computer publicly in the central tax authority office in Lahore at the start of each fiscal year.[11] Apendix Table 10 compares treatment circles to control circles on key tax recovery variables at baseline and shows that they appear balanced.

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

$$\log y_{ct} = \alpha_t + \gamma_t \log y_{c0} + \beta TREAT_c + \epsilon_{ct} \tag{7}$$

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 (7) 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 (7) can be interpreted as the pure incentive effects of the scheme, before any allocations have taken place. We estimate 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). 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

[9] 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 performance was judged using "gross demand" (which is the taxes assessed before exemptions are taken into account).

[10] Note that an additional 3 divisions were added to the randomization in year 2, representing an additional 115 circles and 11 groups, for a total of 525 circles and 52 groups.

[11] 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, groups were not reassigned, and the lottery to assign to assign groups to treatment was conducted province-wide on August 5, 2014.

elicited from inspectors, and the simulations described in Section (3.2), to define the marginal return to effort for each inspector i, $\frac{dE[u_i]}{de_i}$.

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 equation:

$$\begin{aligned} \log y_{ct} &= \alpha_t + \alpha_g + \gamma_t \log y_{c0} + \\ &\quad + \beta_1 TREAT_c \times \frac{dEu_c}{de_c} + \beta_1 \frac{dEu_c}{de_c} + \epsilon_{ct} \end{aligned} \tag{8}$$

where $\frac{dEu_c}{de_c}$ is the estimated marginal utility from effort for inspector c calculated from the baseline preference and predicted y data using equation (4). We include group fixed effects (α_g); since randomization occured by group, this subsumes the main effect of treatment ($TREAT_c$).[12] 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.

5 Results

5.1 Effect of the first year of treatment

We first examine whether the 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 (7) 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. 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 4.8 log points higher than the control group. Compared with the control group's average growth rate of 14.7%, this represents a 44% higher growth rate than controls. For inspectors who were first included in the scheme in the second year the impact is even greater – 8.2 log points higher revenue, or about 80% higher growth rate than controls.

[12] Results without group fixed effects are very similar, but slightly noisier; see Appendix Table 14.

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 diffference 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 of the scheme for inspectors who were included in the scheme in the second year. It is also worth emphasizing that these are purely incentive effects based on expected *future* transfers – 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.

Table 4 disaggregates the results separately 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. While in the first year there is no difference in tax revenue across the two schemes, in year 2 – perhaps as people understood the details of the schemes more – basing the rank-ordering on revenue led to substantially greater increases in revenue than basing the rank-ordering based on the tax base. This is surprising in that even the scheme that rewarded performance on tax revenue collected, as opposed to the tax base, saw a greater impact on both collections and tax base, even though the latter outcome was directly rewarded in the other scheme. While this may sound somewhat counter-intuitive, it likely reflects the fact that there is greater uncertainty about the tax base due to the non-systematic use of exemptions which are meant to be excluded from the net tax base.

The magnitudes of these effects are substantial. By way of comparison, the financial incentive schemes we studied in the same property tax context in Khan, Khwaja, 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 – the pure piece rate scheme – increased revenue by 12.9 log points. The performance-ranked serial dictatorship, studied here, increased total tax revenue by 8.4 log points – two-thirds as large an effect as the maximally effective financial reward scheme we studied, which paid purely based on revenue collected. Yet, the performance-ranked serial dictatorship was completely free to the government, whereas the financial incentives had the government almost doubling the wages of tax staff. This suggests that leveraging postings for incentive purposes can be an extremely cost-effective way for the government to improve performance.[13]

5.2 Heterogeneity by marginal return to effort

While it is reassuring that the scheme on average leads to positive incentives for inspectors to exert effort and in turn to higher tax collections, note that the theoretical analysis in section 3 implied that not all inspectors face equally strong incentives, given heterogeneity in baseline preferences

[13]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. See Feldstein (1999), Chetty (2009), and the related discussion in Khan, Khwaja, and Olken (2016).

and heterogeneity across circles. We now directly test for this heterogeneity by examining whether those inspectors predicted to face higher incentives under the scheme according to the model (i.e. as computed in Section 3.2) do in fact respond more when randomized into the treatment group..

The results, calculated by estimating equation (8), are presented in Table 5. Following the various assumptions used in the simulations above, Panel A presents the results where $\frac{dEu_c}{de_c}$ 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 $\mathbf{P}$ but not $\mathbf{y}$; 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 $\mathbf{y}$ but they assume that other inspectors' preferences $\mathbf{P}$ are random.

The results show that inspectors do indeed respond to the transfer-based incentive scheme more precisely when predicted to do so. To interpret magnitudes, the standard deviation of $\frac{dEu_c}{de_c}$ is 0.37, so the estimates in column 7 of Table 5, Panel A suggest that one standard deviation higher marginal returns increases the average treatment effect on current-year tax collection by 0.09 log points. This suggests that the treatment effects are being driven by the relatively few inspectors with very high marginal incentives.

Figure 4 shows the heterogeneous treatment effects non-parametrically. To do so, we estimate equation (7) for total tax revenue for 100 equally spaced grid-points from 0 to 1.5 based on $\frac{dE[u_1]}{de_i}$. For each gridpoint, we re-estimate equation (7), weighted observations using a triangle kernel based on distance from the grid point.[14] Figure 4 plots the coefficients β, along with 95% confidence intervals. This is analogous to a Fan (1992) regression, but gridding on a different variable to generate heterogeneous treatment effects. The estiamtes confirm that the treatment effects are driven by those with the highest marginal effects. The heterogeneity in treatment effects is particularly concentrated in year 1.

Comparing across the panels in Table 5, in which marginal incentives are calculated using first only information in $\mathbf{P}$ and then only information in $\mathbf{y}$ to disentangle what inspectors appear to react to, it appears that inspectors are responsive to only the aspects of heterogeneity coming predicted differences in outcomes $\mathbf{y}$, and do not respond at all to heterogeneity coming from differences in $\mathbf{P}$. The strongest results come in Panel D, in which inspectors have full knowledge of predicted differences in outcomes $\mathbf{y}$, but assume that other inspectors' preferences are randomly determined.

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 successfully predict which inspectors respond to the incentives. On the net results imply that that – if one knows the distribution of $\mathbf{P}$ and $\mathbf{y}$ – one can use the model in Section 3 to predict whether a particular service or unit would be a good candidate for application of a performance-ranked serial dictatorship.

[14]Note that we estimate each equation differencing out baseline values (i.e. imposing $\gamma_t = 1$) to ensure that the only difference as we move across the grid is heterogeneous treatment effects, not differential persistence of baseline.

5.3 Dynamic effects from repeated application of the transfer scheme

The previous effects focused on the first year the incentive scheme was in place, before any transfers took place. If, however, the policy was in place every year, its effects could be different. There could be allocation effects, for example – people may work better (or worse) once they have been allocated based on their preferences than if just given incentives in places when assignment was exogenous. Relatedly, there could be (adverse) disruption effects as people may perform differently when they have moved to a new place. 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 explore the dynamic effects from having the scheme be repeated multiple times, as described in Section 4 we re-randomized the scheme at the beginning of year 2. This created 4 groups, as shown in Table 1. To analyze the differential effects, we restrict ourselves to year 2 data, and estimate the following regression:

$$\begin{aligned} \log y_{c2} &= \alpha + \gamma \log y_{c0} + \beta_1 TREAT\_Y1_c \\ &\quad + \beta_2 TREAT\_Y2_c + \beta_3 TREAT\_Y1_c \times TREAT\_Y2_c + \epsilon_{ct} \end{aligned} \tag{9}$$

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 of receiving treatment in year 1 and NOT receiving it in year 2; 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.[15]

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. We should note that thet the difference between β_1 and β_2 captures not only differences in (i) persistance (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) 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 preferences) and (iii) any differences between circle allocation (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

[15] Note while that the coefficient on β_2 should be very similar to the year 2 effect in Table 3, since both estimate the effect of starting 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.

the incentives have been turned off.

Second, a key result is the negative interaction term on β_3, which suggests that experiencing the treatment multiple times is less effective than experiencing it once. There are several factors that may be at play here. First, recall that as part of the design inspectors were allowed to change their preferences before the first round of allocations occured (but after they had found out whethere they been reselected for continuation in the scheme or not). This implies that to the extent that those continuing in the scheme changed their preferences, in year 2 the allocation of circles may have been systematically different from those who were no longer continuing treatment. Table 7 checks whether there is indeed such a possible change by examining the preferences submitted by year 1 inspectors at the end of year 1. The results show 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. 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 that those who experienced the treatment in Year 1 may have been more likely to have been moved compared to those who had not, if the performance-ranked scheme creates more movements than occur as part of the status quo. We examine this in Table 8, looking both at a dummy for whether the inspector present the circle at the mid-point of Year 2 (i.e. after transfers 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 Y1 circles were about 10 percentage points more likely to have experienced a move. 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.[16]

A final explanation for the negative interaction effect is simply discouragement: inspectors may 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 1 year, and they need to work hard again in the second year. While we do not have direct evidence for this,. In any case the results presented here suggest that while this type of incentive scheme can be effective, applying it too often can be counter-productive.

6 Conclusion

Effective state bureaucracies play a central role in facilitating growth and development but governments, especially in developing countries, face many constraints on their ability to provide incentives

[16]The increased disruptions may also have had a direct negative effect on revenues. We explore this in more detail in appendix **??**, 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 substantial negative effect of movements on total revenue – a 39 percent decline overall, or 19 percent if we focus on the cleanest estimates (where year 2 treatments are excluded). While these estimates are borderline statistically significant, they are quite noisy and should be interprested with caution. However, they indeed suggest that movements per se do adversely impact performance.

to the agents hired to perform state functions, the bureaucrats. In an attempt to limit politicians' ability to use government jobs to reward political cronies, many governments have adopted strict civil service rules. In these systems, pay and promotion are often rigid and mechanical, usually based on initial level and seniority rather on performance.

Transfers, or horizontal movements, provide a feasible and promising avenue for rewarding good performers and punishing bad performers, but the ambiguity of assignment rules and issues with revelation of agents' preferences over postings 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 (PRSD), 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 bureacracy in Punjab, Pakistan, that formalizing the relationship between peformance and transfers indeed improves performance. Indeed, by the second year of our study, those tax inspectors randomly allocated to the performance-ranked serial dictatorship had an 80 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 results suggest that bureaucracies have tremendous potential to improve performance at zero cost by periodically using transfers as an incenttive, particularly when preferences over postings have a substantial common component.

References

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

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

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

Fan, J. (1992): "Design-adaptive nonparametric regression," *Journal of the American statistical Association*, 87(420), 998–1004.

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

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

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

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

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

Lazear, E. P., and S. Rosen (1981): "Rank-Order Tournaments as Optimum Labor Contracts," *The Journal of Political Economy*, pp. 841–864.

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

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

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

Table 1: Treatment assignment of circles in year 2

	Year 2 Control	Year 2 Treatment	Total
Year 1 Control	210	56	266
Year 1 Treatment	69	75	144
(Not included in Year 1 lottery)	96	19	115
Total	375	150	525

Table 2: Comparing simulation results to "reduced form" results

	Full knowledge of P, y	Full knowledge of P, no knowledge of y	Assume identical P, full knowledge of y	Assume random P, full knowledge of y
	(1) dEu1dy	(2) dEu1dy	(3) dEu1dy	(4) dEu1dy
Reduced form preference correlation (ρ)	0.476***	0.629***	0.141***	0.205***
Reduced form Y density (γ)	0.021	-0.094*	0.236***	0.067
Outcome stdev	0.559***	0.809***	0.152***	0.274***
Fraction weakly better	-0.415***	-0.355***	-0.225***	-0.283***
Fraction strictly better	0.088*	0.233***	-0.047	-0.043
Fraction weakly worse	-0.047	-0.170***	0.066	0.067
Fraction strictly worse	0.424***	0.369***	0.229***	0.288***

Notes: Correlation coefficient of dEu1dy and various measures of returns to effort. * p<0.10, ** p<0.05, *** p<0.01

Table 3: Treatment Effect on 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	0.048**	0.044*	0.067	0.082**	0.074*	-0.118	0.058***	0.053**	0.012
	(0.023)	(0.024)	(0.058)	(0.041)	(0.038)	(0.118)	(0.020)	(0.021)	(0.054)
N (Total)	405	405	396	259	259	251	664	664	647
Mean of control group (Total)	15.907	15.692	14.072	16.255	16.134	13.805	16.061	15.888	13.954

Notes: OLS regessions of log recovery on treatment assignment. 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. * p<0.10, ** p<0.05, *** p<0.01

Table 4: Treatment Effect on Tax Revenue, by Sub-Treatment

	Year 1			Year 2			Pooled		
	(1) Total	(2) Current	(3) Arrears	(4) Total	(5) Current	(6) Arrears	(7) Total	(8) Current	(9) Arrears
Panel A: Effect on Recovery									
Revenue	0.058	0.041	0.041	0.131***	0.116**	-0.147	0.080***	0.064**	-0.019
	(0.036)	(0.035)	(0.083)	(0.048)	(0.052)	(0.129)	(0.029)	(0.030)	(0.070)
Demand	0.037*	0.046	0.097	0.005	0.008	-0.075	0.031	0.039	0.049
	(0.022)	(0.031)	(0.066)	(0.061)	(0.044)	(0.204)	(0.023)	(0.026)	(0.071)
N (Total)	405	405	396	259	259	251	664	664	647
Mean of control group (Total)	15.907	15.692	14.072	16.255	16.134	13.805	16.061	15.888	13.954
Revenue = Demand (p-value)	0.586	0.914	0.563	0.086	0.096	0.757	0.152	0.510	0.462
Panel B: Effect on Net Demand									
Revenue	0.110**	0.064	0.188	0.109	0.069	-0.226	0.109**	0.065*	0.079
	(0.054)	(0.047)	(0.134)	(0.067)	(0.060)	(0.225)	(0.044)	(0.038)	(0.117)
Demand	0.022	0.012	0.114	-0.008	-0.012	0.146	0.016	0.007	0.108
	(0.039)	(0.029)	(0.116)	(0.091)	(0.051)	(0.320)	(0.035)	(0.026)	(0.111)
N (Total)	406	405	388	204	204	196	610	609	584
Mean of control group (Total)	16.411	16.317	13.854	16.605	16.471	14.129	16.485	16.376	13.960
Revenue = Demand (p-value)	0.136	0.314	0.646	0.272	0.254	0.325	0.066	0.181	0.843

Notes: OLS regessions of log net demand on treatment assignment. Note that Net Demand outcomes are measured using values from the first quarter of the following fiscal year. The unit of observation is a circle, as defined at the time of randomization. Specification controls for baseline value. Robust standard errors in parentheses. Standard errors are clustered by circle. * p<0.10, ** p<0.05, *** p<0.01

Table 5: Heterogeneity in treatment effects by simulated marginal returns to effort

	Y1 Q4			Y2 Q4			Pooled		
	(1) Total	(2) Current	(3) Arrears	(4) Total	(5) Current	(6) Arrears	(7) Total	(8) Current	(9) Arrears
Panel A: Full knowledge of P, y									
Treatment * dEu1dy	0.351**	0.291*	0.284	0.166**	0.149*	0.142	0.250***	0.211***	0.349
	(0.160)	(0.154)	(0.229)	(0.068)	(0.078)	(0.431)	(0.093)	(0.074)	(0.232)
dEu1dy	0.025	0.133*	-0.145	-0.077	0.005	-0.252**	-0.016	0.084	-0.217***
	(0.066)	(0.075)	(0.105)	(0.061)	(0.071)	(0.123)	(0.047)	(0.050)	(0.074)
N	403	403	394	257	257	249	660	660	643
Mean of control group	15.910	15.698	14.069	16.261	16.141	13.804	16.066	15.893	13.952
Panel B: Full knowledge of P, no knowledge of y									
Treatment * dEu1dy	0.062	0.001	0.005	0.007	-0.046	0.599	0.050	-0.006	0.285
	(0.069)	(0.066)	(0.235)	(0.110)	(0.074)	(0.525)	(0.055)	(0.045)	(0.177)
dEu1dy	-0.025	0.055	-0.051	-0.052	0.063	-0.431***	-0.040	0.055*	-0.232***
	(0.042)	(0.051)	(0.114)	(0.042)	(0.046)	(0.129)	(0.029)	(0.031)	(0.077)
N	403	403	394	257	257	249	660	660	643
Mean of control group	15.910	15.698	14.069	16.261	16.141	13.804	16.066	15.893	13.952
Panel C: Assume identical P, full knowledge of y									
Treatment * dEu1dy	0.192**	0.099	0.378**	0.113	0.211	-0.222	0.177***	0.162**	0.370*
	(0.081)	(0.098)	(0.178)	(0.142)	(0.166)	(0.434)	(0.066)	(0.080)	(0.185)
dEu1dy	0.097	0.268**	-0.205	0.076	0.042	0.117	0.082	0.179*	-0.158
	(0.089)	(0.111)	(0.126)	(0.128)	(0.158)	(0.201)	(0.073)	(0.095)	(0.117)
N	403	403	394	257	257	249	660	660	643
Mean of control group	15.910	15.698	14.069	16.261	16.141	13.804	16.066	15.893	13.952
Panel D: Assume random P, full knowledge of y									
Treatment * dEu1dy	0.764***	0.694**	0.820	0.468***	0.504**	-0.224	0.582***	0.563***	0.783*
	(0.244)	(0.267)	(0.561)	(0.157)	(0.184)	(0.822)	(0.150)	(0.144)	(0.427)
dEu1dy	0.141	0.395**	-0.321	-0.065	-0.038	-0.082	0.084	0.266	-0.293
	(0.161)	(0.192)	(0.247)	(0.185)	(0.209)	(0.240)	(0.136)	(0.160)	(0.179)
N	403	403	394	257	257	249	660	660	643
Mean of control group	15.910	15.698	14.069	16.261	16.141	13.804	16.066	15.893	13.952

Notes: OLS regessions of log recovery on treatment assignment, with group fixed effects. The unit of observation is a circle, as defined at the time of randomization. Specification controls for baseline value. Robust standard errors in parentheses. Standard errors are clustered by circle. * p<0.10, ** p<0.05, *** p<0.01

Table 6: Dynamic effects estimated in year 2

	(1) Total	(2) Current	(3) Arrears
Y2 Treatment	0.069*	0.058	-0.121
	(0.041)	(0.040)	(0.118)
Y1 Treatment	0.108***	0.089**	0.139
	(0.040)	(0.042)	(0.101)
Y1 AND Y2 Treatment	-0.133**	-0.093	-0.004
	(0.068)	(0.069)	(0.179)
N (Total)	403	403	392
Y1 Treatment = Y2 Treatment (p-value)	0.464	0.575	0.070
Mean of control group (Total)	16.255	16.134	13.805

Notes: OLS regessions of log recovery 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. * p<0.10, ** p<0.05, *** p<0.01

Table 7: Do preferences depend on continuing status?

	(1) Own circle is favorite	(2) Own circle is favorite	(3) Own circle u_1	(4) Own circle u_1
Continuing	0.140	0.146*	0.038	0.031
	(0.085)	(0.079)	(0.051)	(0.047)
Own circle is favorite, baseline		0.344***		
		(0.083)		
Own circle u1, baseline				0.389***
				(0.115)
N (Total)	108	107	108	107
Mean of non-continuing group (Total)	0.660	0.660	0.854	0.854

Notes: OLS regessions on continuing treatment assignment. Sample is restricted to Y1 treatment inspectors only. The unit of observation is an inspector. Robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

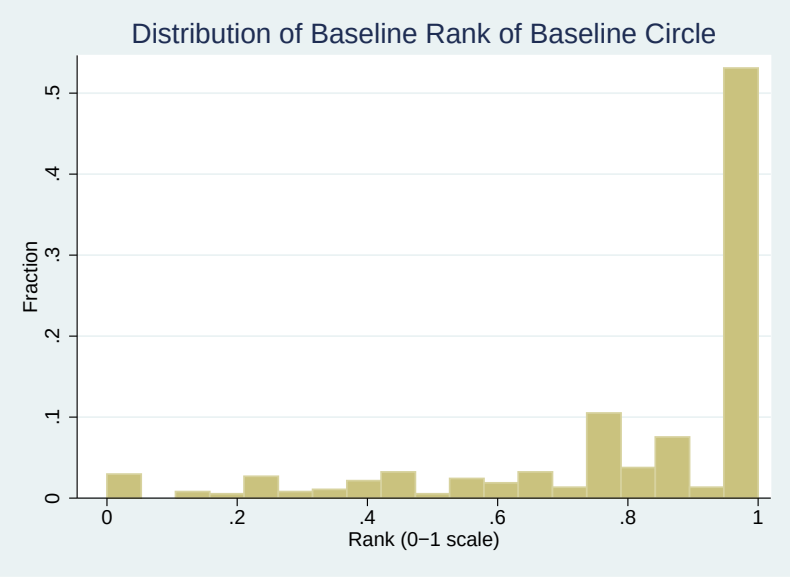
Table 8: How does the serial dictatorship change allocations?

	(1) Any move	(2) Any move	(3) Days in circle	(4) Days in circle
Y2 Treatment	-0.044		12.290	
	(0.082)		(40.755)	
Y1 Treatment	0.101	0.092*	-65.624*	-56.809**
	(0.071)	(0.054)	(36.251)	(27.127)
Y1 AND Y2 Treatment	0.010		9.343	
	(0.118)		(59.590)	
N (Total)	365	365	365	365
Y1 Treatment = Y2 Treatment (p-value)	0.129		0.106	
Mean of control group (Total)	0.555	0.546	388.476	390.903

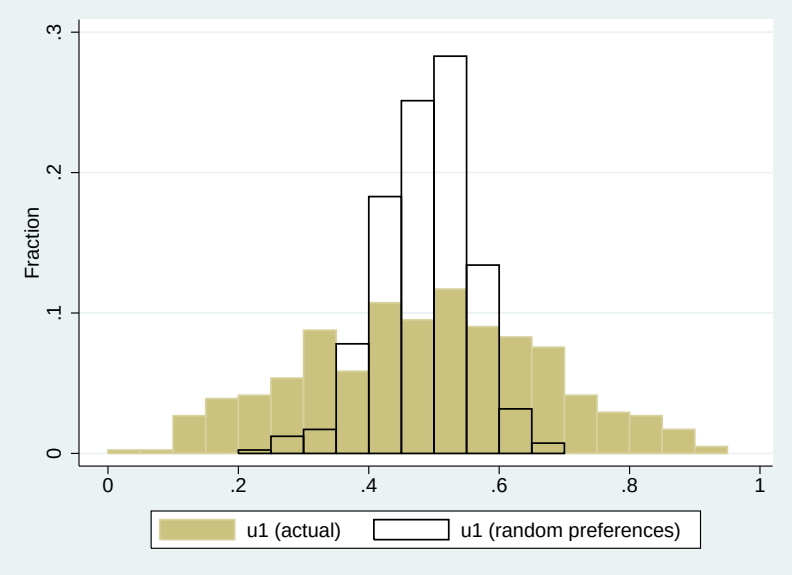
Notes: OLS regessions of number of days in circle or dummy for any move on various treatment regressors. LHS variables are calculated over the time period from the beginning of FY2014 to the date of the execution of transfers. The unit of observation is a circle, as defined at the time of randomization. Sample excludes any circles that have been merged or split after ballot. Specification controls for baseline values. Robust standard errors in parentheses. Standard errors are clustered by circle. * p<0.10, ** p<0.05, *** p<0.01

Figure 1: Descriptive statistics of baseline preferences over positions

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



(b) Distribution of average circle ranks



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 histogram of average ranks of a circle j, averaged over all inspectors in the group. Ranks are normalized (prior to averaging) such that the top-ranked circle is normalized to 1 and the bottom ranked circle is normalized to 0. The histogram in outline shows what the distribution would look like if inspectors' preferences were random.

Figure 2: Example of marginal incentives with different error variances

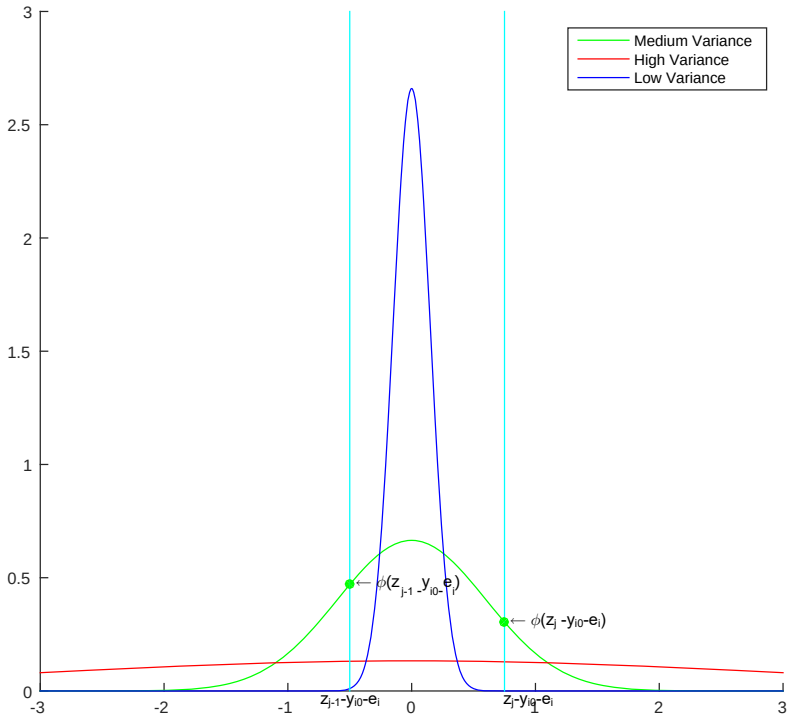
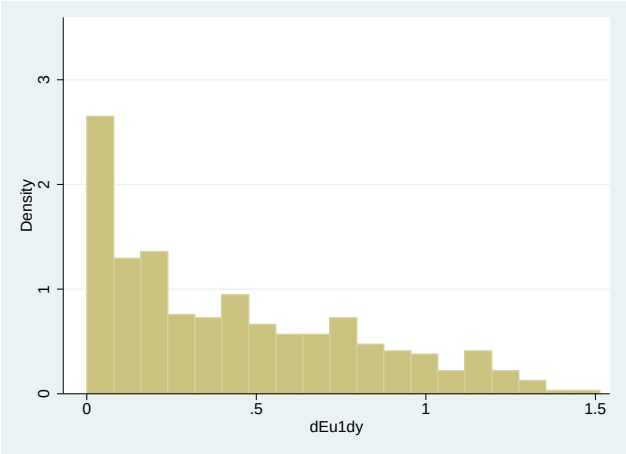
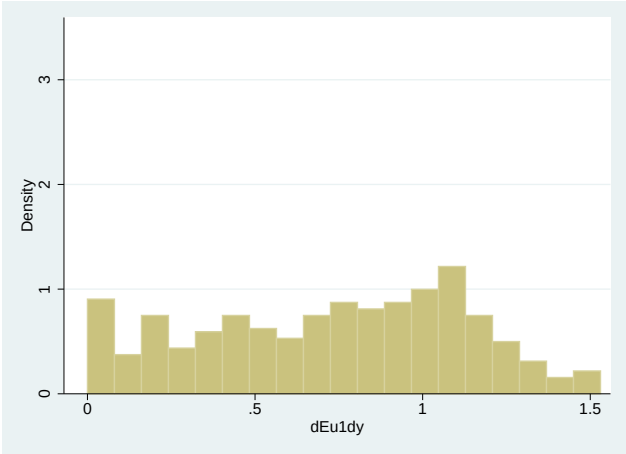


Figure 3: Simulated Distribution of $\frac{dE[u_1]}{de_i}$ under alternative assumptions about knowledge

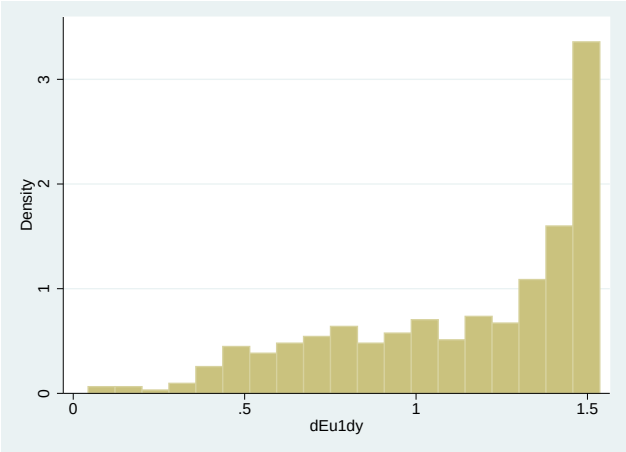
(a) Full Knowledge of P and y



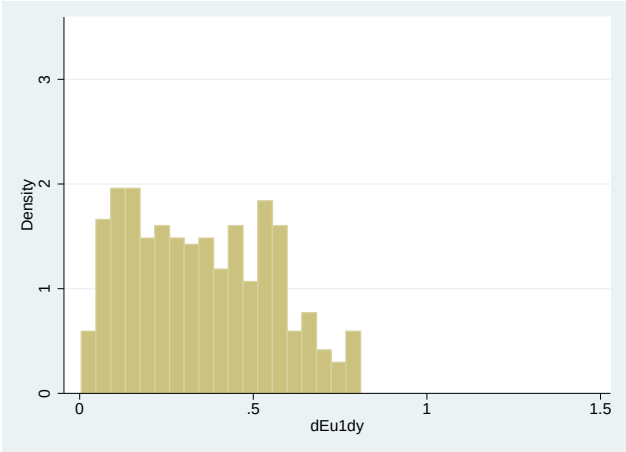
(b) Full knowledge of P, no knowledge of y



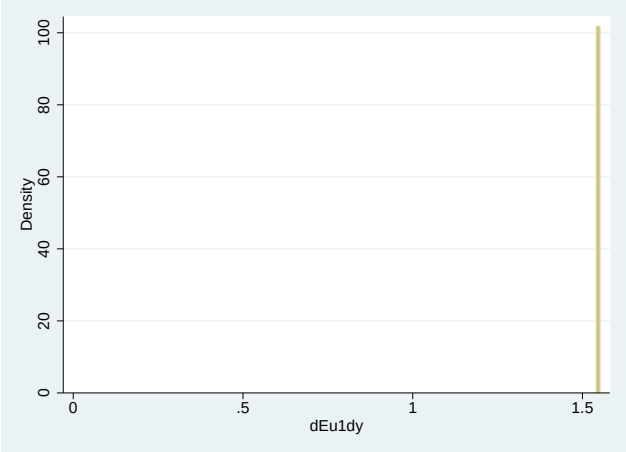
(c) Assuming identical preferences P, full knowledge of y



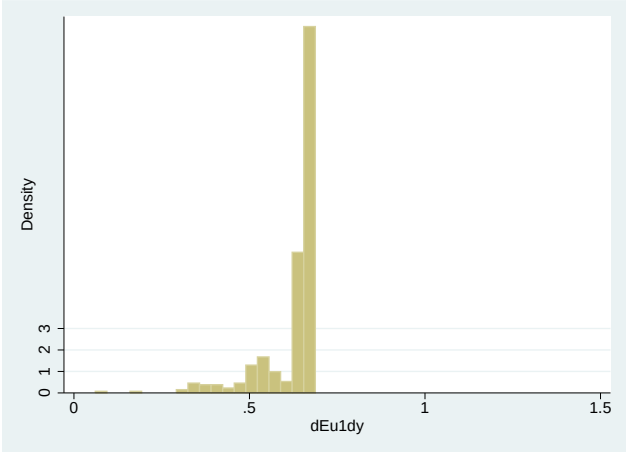
(d) Assuming random preferences P, full knowledge of y



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



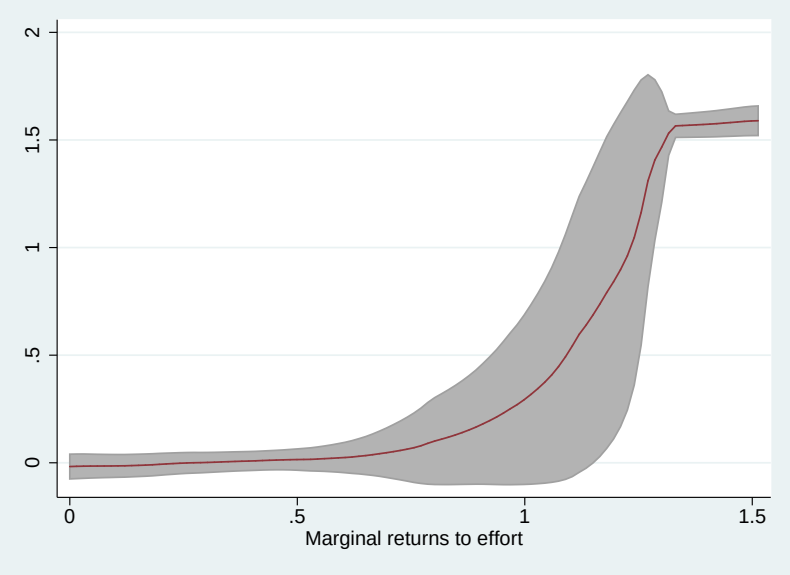
(f) Assuming random preferences P, no knowledge of y



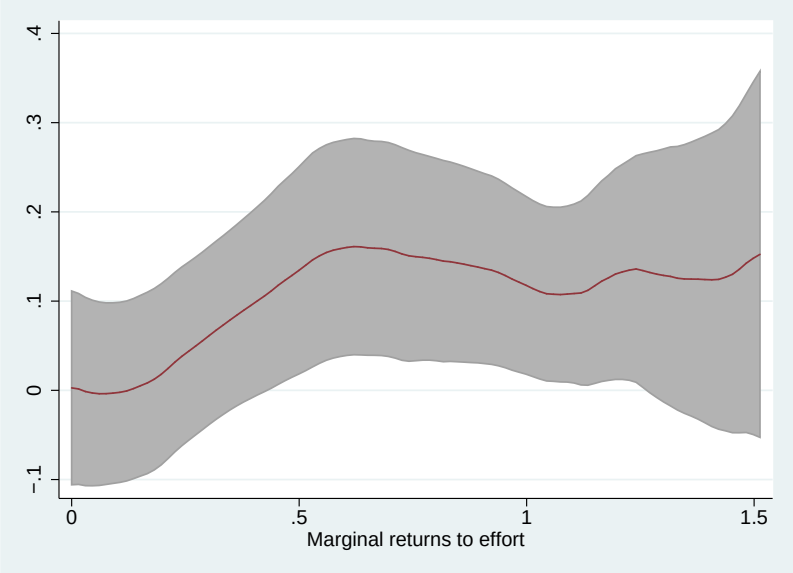
Notes: Simulations are as described in Section (3.2).

Figure 4: Non-parametric Heterogeneous Treatment Effects by $\frac{dE[u_1]}{de_i}$

(a) Year 1



(b) Year 2



Notes: Figure shows estimated effects β from a series of estiamtions of equation (7). We estimate equation (7) for 100 equally spaced grid-points from 0 to 1.5 of $\frac{dE[u_1]}{de_i}$; for each estimation, observations are weighted using a triangle kernel based on distance from the grid point. 95% confidence intervals are plotted. We estimate each equation differencing out baseline values (i.e. imposing $\gamma_t = 1$) to ensure that the only difference as we move across the grid is heterogeneous treatment effects, not differential persistence of baseline. Plotted results are for log total tax revenue.