Should Governments Measure What They Manage? Experimental Evidence from Monitoring Bureaucrats in India

Advitha Arun[*] Steven Brownstone[†] Siddharth George[‡]
Naveen Kumar[§] Karthik Muralidharan[¶]

27 April 2025

Abstract

Bureaucrats implement most development programs and collect administrative data on the programs they manage. Monitoring this data can facilitate learning and performance improvement but creates incentives to misreport. We experimentally study the impacts of providing bureaucrats with performance scorecards based on their self-reported data. Scorecards helped bureaucrats become better informed about their performance. Nevertheless, scorecards on average resulted in slightly worse public service delivery. Using audit-based measures of data integrity, we find evidence that scorecards caused some bureaucrats to fudge admin data to inflate their scores. This effect overshadowed the positive effects we observed for bureaucrats who learned the most. Documenting the heterogeneous treatment effects using generalized machine learning further confirms that there are sizable groups of bureaucrats where the scorecards succeeded despite the negative effects overall.

[*]UC Irvine, Email: advitha.arun@uci.edu
[†]UC San Diego, Email: sbrownst@ucsd.edu
[‡]National University of Singapore, Email: segeorge@nus.edu.sg
[§]BGS College of Engineering & Technology, India, Email: nkumar1@bgscet.ac.in
[¶]UC San Diego, Email: kamurali@ucsd.edu

1 Introduction

Most development programs are ultimately implemented at scale by bureaucrats. The effectiveness of development schemes thus depends critically on the performance of public officials like city managers, school teachers, primary health workers, tax collectors, and agricultural extension officers. Public sector organisations can improve bureaucrat performance and service delivery by adopting effective personnel policies and management practices (Rasul and Rogger, 2013). One common strategy, particularly in contexts where performance-based financial incentives are infeasible, is to monitor bureaucrat performance with dashboards not explicitly tied to high-powered incentives (Finan, Olken, and Pande, 2017).

The impact of dashboards is unclear. They can help bureaucrats learn how to improve at their jobs and enable the use of non-pecuniary incentives like postings and awards (Khan, Khwaja, and Olken, 2019). However, dashboards can also cause bureaucrats to focus on metrics over performance. In many cases, bureaucrats have some ability to manipulate the data used to monitor them, and even without manipulation, focusing only on tasks captured in monitoring dashboards can lead bureaucrats to ignore other services valued by citizens.

Dashboards without high-powered incentives and some degree of self-reporting are the most common form of performance management intervention in bureaucracies. Prior research focuses on cases where high-quality third-party data is available to form the basis of often higher-powered performance management systems (Besley, Burgess, Khan, and Xu, 2022; Finan, Olken, and Pande, 2017). In this paper, we examine the impact of monitoring the performance of Gram Panchayat Secretaries (PSes) in the Indian state of Telangana. Gram Panchayats (GPs) are the lowest tier of government in rural India, responsible for delivering local public services and running government schemes.

Each GP is managed by a PS, who is in charge of day-to-day operations and administration. PSes hire local staff, manage public infrastructure (e.g. roads, drains, streetlights, plantations), oversee public services (e.g. sanitation), interact with citizens (e.g. registering births and deaths), and implement key development programs (e.g. the National Rural Employment Guarantee Scheme). India has approximately 250,000 GPs serving more than 900 million citizens, each administered by a PS. Despite their prevalence and importance, which has been highlighted in qualitative research (Rao, Ananthpur, and Malik, 2017; Veeraraghavan, 2021) and even popular culture (e.g. this popular comedy-drama), PSes remain virtually unstudied by economists.

We find evidence that on average the dashboards modestly decreased service delivery, as measured from citizen surveys, and increased misreporting, measured by comparing bureaucrat and supervisor reports. This average effects mask substantial individual heterogeneity. Using modern ML methods to construct individual treatment effects, we estimate that the dashboards had a positive impact on 34% of the bureaucrats studied. While, the positive effect is weakly related to the degree of learning, the individual treatment effects themselves are not significantly correlated with any specific observable covariate.

Measuring the performance of a PS is challenging, given the breadth of the job. We worked with the state to design and introduce a mobile application for PSes to capture or report data across all of their responsibilities[1]. Using the self-reported data from the performance management app we determine a performance score for each PS. The score is a weighted average of performance across more than ten responsibility areas. We experimentally evaluate the impact of giving performance scorecards generated using self-reported data. 75% of all PSes were randomly selected to receive monthly scorecards.

With self-reported data, governments face a trade-off between using the data for performance management and trusting the integrity of data (Singh, 2020)[2]. We evaluate the extent of this trade off between learning and misreporting by collecting the same self-reported data from alternative sources. First, the mandal supervisors (MPOs) periodically visit villages under their supervision and collect very similar data to the PSes daily reports. We deliberately aligned these reports so that we could construct a measure of "truthful" reporting for the PSes on the days where their supervisor came for an inspection. Even more rarely, state level visitors (SLVs) visit villages to collect data on a subset of the outcomes covered by the PS and their supervisor. Finally, the department has a call center which periodically conducts phone surveys of citizen satisfaction. The SLV and citizen data allow us to validate the supervisor data used in the main analysis.

We explore three reasons why scorecards did not improve bureaucrat performance in our context. First, we show that while scorecards helped bureaucrats learn about the important aspects of their jobs and their own weak areas, this learning did not lead to improved performance. Second, we show that scorecards had similar effects on the performance of tenured and untenured bureaucrats, suggesting that the availability and salience of per-

[1]Not all performance indicators can be captured by devices, but instead will have to be self-reported by the bureaucrats. Like many apps for frontline workers photos and capturing GPS locations tags were the primary methods to ensure data credibility.

[2]Collection of outcome data can be delegated to third parties to ensure reliability but may limit external validity as bureaucrats are usually tasked with collecting administrative data on the programs that they administer.

formance data did not heighten the impact of career concerns in our setting. Third, we show that while scorecards enabled the use of social incentives such as public recognition for good performance, this had no impact on actual performance. Thus, mechanisms through which monitoring has been shown to improve bureaucrat performance in other settings did not operate in our context.

We also present evidence that scorecards increased multi-tasking problems. Scorecards reduced citizens' overall evaluations of public service delivery, We find that scorecards improve evaluations for high-weightage indicators, while reducing evaluations for low-weight indicators. Thus, we present evidence for two types of multi-tasking problems — increased misreporting and the more traditional multi-tasking problem of greater effort spent on monitored indicators.

In the last part of the paper, we conduct a simple meta-analysis to compare our results against those of other monitoring interventions. We find that very few monitoring interventions focus on coping jobs — ours is, to our knowledge, the only experimental evaluation of performance monitoring for a job approximating the core civil service. We also observe that performance monitoring tends to have less positive effects when bureaucrats have greater potential for multi-tasking and when they are responsible for collecting the data that is used for performance evaluation.

Overall, our results highlight conditions under which performance monitoring is most likely to be beneficial and improve bureaucrat performance.

Contribution to Literature. Our paper contributes to two strands of literature. First, we contribute to a growing literature on the personnel economics of performance measurement and monitoring (See (Besley, Burgess, Khan, and Xu, 2022; Finan, Olken, and Pande, 2017) for a review). Examples of measurement include tax collectors (Khan, Khwaja, and Olken, 2016, 2019), teachers (Duflo, Hanna, and Ryan, 2012), agricultural extension workers (Muralidharan, Niehaus, Sukhtankar, and Weaver, 2021) and officials tasked with road construction (Olken, 2007). We contribute to this literature by evaluating the impacts of performance monitoring for jobs that approximate the core civil service. In these multi-faceted jobs, monitoring may have more positive impacts in terms of fostering learning, but also create more opportunities for multi-tasking and manipulation. We show that monitoring has less sanguine effects in our setting, given the opportunity to misreport data and inflate performance (Jacob and Levitt, 2003; **?**).

Second, we contribute to the literature on multi-tasking. Prior work has shown that

high-powered incentives in government can backfire (Acemoglu, Fergusson, Robinson, Romero, and Vargas, 2020; Chen, Li, and Lu, 2018; de Janvry, He, Sadoulet, Wang, and Zhang, 2020; Fisman and Wang, 2017; Giné, Mansuri, and Shrestha, 2022). We provide evidence of two types of multi-tasking. We show that PSes respond to scorecards by misreporting data and inflating their performance. We also find that PSes improve performance on high-weight indicators while worsening performance on low-weight indicators, including those that citizens value.

The first part of the paper describes the institutional setting and treatments. Then, we discuss the various data sources used and constructed for the study. We offer some initial correlations suggesting that the administrative scores do correlate with citizen opinions. We then show treatment effects on the different performance measures. Finally, we look at heterogeneous treatment effects both traditional and using machine learning. Finally, we discuss the mechanisms suggested by the results and directions for future research.

2 Institutional Setting

In this section, we describe the policy that led to the hiring of 9,355 Panchayat Secretaries to implement implement all public works realted to the village assigned. We further describe the roles and responsibilities of PSes, the performance measurement system, and inspections by PSes supervisors. In particular, we explain the nature of the inspections and and practical challenges that come with PSes job.

2.1 Hiring of Panchayat Secretaries

A state is administered through districts and each district is broken down into mandals (blocks). Each mandal is then a made up of several gram panchayats. Gram panchayat is the grass root level of administration in India. It provides a democratic structure for villages through elected members. Each village with a population of 5,000 or a group of villages get to form a gram panchayat (GP).

Telangana region in the unified state of Andhra Pradesh was given statehood in 2015. Government of Telangana implemented the "Panchayat Raj" Act in 2018. To monitor and support a gram panchayat, the act allows the state to appoint panchayat secretaries to all gram panchayats. The act also created a number of new Gram Panchayats to bring local administration closer to citizens. Telangana has a population of 35 million and with 33 districts, 594 mandals, and 12,769 GPs.[3]

To ensure that each GP has a PS, the GoT allowed Commissioner Panchayat Raj & Rural Employment to fill 9,355 vacant posts of Junior Panchayat Secretaries (JPS) in 2018. The JPSes are hired on a remuneration basis for 3 years and they were to be regularised as permanent government employees based on satisfactory performance. The JPSes join existing panchayat secretaries who are already working as frontline bureaucrats. When the study started in April 2021, the state of Telangana had 12,101 panchayat secretaries for 12,769 GPs.[4] Of the 12,769 GPS, JPSes manage 9,081 GPs and PSes manage 2,839 GPs.[5]

[3] At the time of formation, Telangana had 10 districts, 459 mandals and 8,368 gram panchayats.

[4] 668 GPs were given as "full additional charge" or "incharge".

[5] The remaining 849 GPs are either managed by a contract PS or "Outsourcing" PS which have even less job security than junior panchayat secretaries

2.2 Roles and Responsibilities of PSes

According to the Act, PSes are responsible for all public works related to the GP assigned. The panchayat secretaries manage small teams of laborers to clean roads and drains, maintain streetlights, plant greenery, produce compost, and collect garbage. The secretaries are also responsible for managing village finances, collecting property taxes, and providing various certificates and approvals such as birth and death records. The act also defines Key Result Areas (KRA) that are a priority to the government.

Prioritizing key result areas largely takes place at the state level. At the time of the app release, sanitation is by far the most important result area to the administration followed by plantations. The Act specifies specific targets such as 90% working streetlights. PSes have several general responsibilities in governance, administration, and tax collection, implementation of development schemes. The administration values general administrative result roughly the same as more measurable result areas.

2.3 Introduction of the Performance Measurement System

Unlike tax collectors, but similar to other multifaceted government jobs, capturing performance metrics is difficult. The panchayat secretaries may themselves not understand if they are performing well or poorly. We worked with the government of Telangana to design and introduced an Android mobile application for panchayat secretaries to collect and report data across all of their areas of responsibility.

The mobile application captured daily information on the status of randomly selected roads and drains within the GP. On a monthly basis, the application captured information on GP governance and other aspects of GP performance. Like many apps for frontline workers, photos and capturing GPS locations tags were the primary methods to ensure data credibility. Starting December 2020, the commissioner of panchayat raj instructed all panchayat secretaries of the state to use the mobile application for the reporting their daily, weekly and monthly responsibilities.

2.4 GP Inspections

All panchayat secretaries within a mandal are supervised by the Mandal Panchyat Officer (MPO). Recall that the Panchayat Raj Act 2018 states JPSes will be hired on probationary basis for 3 years and they would be regularised based on satisfactory performance. The

Panchayat Raj Act 2018 also instructs the MPOs to conduct GP inspections. This was the administration's plan to gather the performance indicators needed for JPSes regularisation decision. Importantly, there is substantial heterogeniety in the number of GPs per mandal. MPOs can manage anywhere between 3 and 32 PSes generating substantial variation in the intensity of supervision and frequency of inspection.

All MPOs within a district in turn report to the District Panchayat Officer (DPOs). The commissioner of Panchayat Raj, through DPOs, implements and supervises the administration's efforts to supports and improve various aspects of GPs. The administration had also rolled out a system of state level visitors (SLVs) who periodically inspected GP to assess them. These visitors are generally retired officials which were specifically recruited for the purpose of auditing local projects. There data serves as the external performance indicators of GPs and came to be particularly valued by senior state leadership for tracking progress. However, this external data cannot be used in JPSes regularisation decision as the Act requires that only the Panchayat Raj Departments data (such as MPO inspections) shall be considered for regularisation decision. Also, with only 103 state level visitors and other inspection responsibilities these state level visitors (SLVs) could only cover approximately 2,000 GPs per month which means the data couldn't be used to track GP performance at high frequency.

2.5 Practical Challenges that come with PSes Job

Similar to how a state has a chief minister elected by the people and a bureaucrat (IAS) appointed by the state, a GP has a sarpanch (elected) and a PS (appointed). Sarpanchs and Panchayat Secretaries work very closely together for all GP related activities. However, their mixture of cooperation and conflict is described by some officials as akin to a wife and mother-in-law in a household. Aside from controlling the budget, the sarpanchs influence contracting and hiring for basic public works. There are a key village resources that sarpanch's can co-opt for example access to a tractor which is necessary for successful plantations. Similarly, having a computer in the GP is a key piece of enabling infrastructure.

There are a variety of socio-economic factors that affect PSs ability to achieve the key results enumerated in the Panchayat Raj act. GPs with large SC/ST populations can be more difficult to administer. For example, village sanitation personnel (even if they are SC/ST) are often unwilling to clean the SC/ST hamlet because it is beneath them. The physical size of a GP is important. The funding doesn't really take into account the physi-

cal spread of a GP and it can be physically impossible to clean larger GPs with the allotted funds. Finally, junior panchayat secretaries' own lack of training makes achieving the results envisioned in the act difficult. The government has not carried out a systematic training program for the JPSs.

The incentive for JPSes to perform well at job is clear - permanent employment. For panchayat secretaries on permanent contracts transfers, remain a powerful to reward or punish panchayat secretaries. Since the JPSs are younger and have less training than can be more easily bossed around by the Sarpanchs. The more permanent PSs are sometimes more powerful than the sarpanchs since they have more education. The permanent PSes primary incentive is transfers to larger, more resourced, or more welcoming GPs. As mentioned above there is substantial heterogeniety in the difficult of the PS job across GPs. Some GPs have substantial tax revenue which makes the PSes job easier since they can hire more staff. PSes cannot be directly promoted to MPOs. They must take another competitive examination to be eligible for the MPO job.

3 Treatment

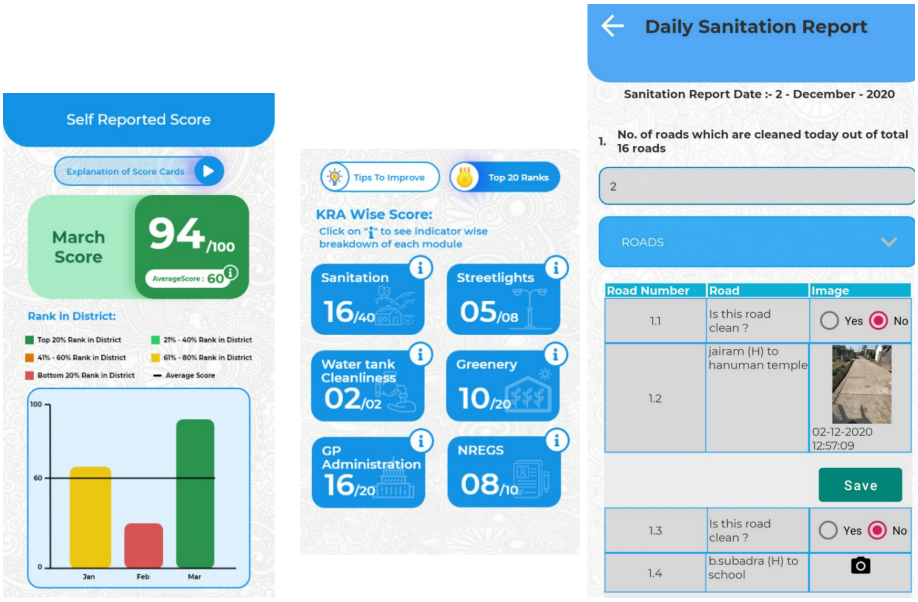
3.1 Creating Scorecards

Using the self-reported data from the app, we determine a performance score for each GP and assign that score to the PS. The score, which is out of 100, is a weighted average of performance across more than ten responsibility areas. Recall that the PS is responsible for all GP related tasks. Table x lists the composition of the overall score. The 100 points is distributed across y key result areas as follows: 40 points for sanitation, 8 points for streetlights, 2 points for water tank, 20 points for greenery, 20 points for GP administration and 10 points for NREGS. See appendix C for the exact rubric.

Table 1: GP performance/progress indicators - change this to a figure

	Key Result Areas (PR Act)	Notes	Points
1	Sanitation	Road, Drain, and Institutional Cleanliness	40
2	Street lights	Installation, maintenance, performance	8
3	Water Tanks	Water tank cleanliness	2
4	Plantations	Nursery, plantations, survival	20
5	Administration	Meetings, Certificates, and Registers	12
6	NREGA	Days worked per active job card	18

Figure 1: Screenshots from PS App



3.2 Sharing Scorecards

(J)PSes of GPs in the treatment group received additional screens in their mobile applications which displayed the GP's overall score and rank based on the information PSes had entered over the course of the month Figure 1 [6]. This setting allows us to experimentally evaluate the effect of providing PSes performance-based feedback via app-based scorecards that show the PS her absolute level of performance, highlight areas for improvement, and convey relative performance within the district.

We hypothesize that the scorecards might improve true performance through providing information that helps PSes learn how to do better as well as by providing incentives to perform well and stand out as a top performer (or at least not stand out as a bottom performer). The scorecards in the app also displayed panchayat secretaries relative rank band within their district and the names of the top 20 panchayat secretaries in the district.

3.3 Validating Self-Reported Data Through MPO Inspections

We examine whether performance-based feedback actually improves true performance – or merely leads to misreporting. Misreporting is when (J)PSes fudge self-reported perfor-

[6] While the value of each indicator was clearly displayed within the application itself, the actual method of score computation was a separate PDF document that panchayat secretaries had to access. Consequently, not all Panchayat secretaries were immediately aware that their own data was being used to generate the scores.

mance data to score better on a metric that now has higher stakes. We leverage on MPO inspections to capture misreporting. In a mobile app designed for inspections, we also ask MPOs to enter data on all PS indicators that can be compared/verified. Additionally, we also modified state level visitors inspection app to include the same indicators to use as external audit. However, since the external visitor app was built by a different technology contractor many of the crucial sanitation questions which relied on road level answers couldn't be used for direct comparison.

4 Randomization and Timeline

The mobile application for the PS to self-report on duties was launched in December 2020. It took some time to fix bugs and drive usage from January to March 2021. By April 2021 we were able to generate scores from the app. However, the department did not share a scorecard for April since that corresponded with India's intense but brief delta COVID wave. The first scorecard was released for May on June 27th. Since some indicators are entered at the end of the month we still consider June part of the treatment period, but July was the first full month where treatment PSes were able to react to scorecards. The July, August, and September scorecards were released closer to the end of their respective months. In October, there was a change in leadership of the Panchayat Raj department. The new leadership wanted a large number of changes in the application which resulted in data collection temporarily pausing. The app, with some significant modifications focusing on PS attendance, relaunched in November 2021. The department was about to relaunch the scorecards but then ran into contracting issues with the contractor for the mobile application. The mobile application stopped working again in May 2022 and was recently restarted under a new tech contractor in September 2022.

We conducted a stratified randomization at the mandal level using district and number of GPs per mandal as strata. 9,584 GPs in 405 mandals received scorecards and 3,186 GPs in 134 mandals didn't receive scorecards. We didn't stratify on baseline scores since the department required we randomize prior to baseline scores becoming available. Further, since randomization took place at the mandal level averages of GP level characteristics were less useful. Mandal level randomization was necessary since PSes within a mandal meet frequently and providing scorecards to some PSes and not others would sow confusion and distort supervisor effort. The mandal size strata was chosen since we believed the intensity of supervisory vists would strongly influence PSes response to scorecards. We chose a split of 75% and 25% control based on plans to layer on further cross-randomized treatments. However, these plans were disrupted by the October leadership change discussed above.

5 Data

4.1 Validation with Citizen Surveys

To validate the performance scores we conducted citizen phone surveys with numbers sampled from government databases. In total we sampled 26,775 citizens across 3,825 GPs. In these phone surveys we asked citizens to rate how their GP was performing

on different result areas themselves using a simple five point likert scale. Table 2 shows that the PS scores, pooled across all months, correlated significantly with the citizen feedback. The scores awarded by inspectors (SLVs as well as MPOs) also correlates significantly with the citizen satisfaction score. This suggests that the inspections and the scores awarded by inspectors are a signal of PS performance on the ground.

Table 2: Correlation Matrix: PS Performance Measures

	(1)			
	PS Score	MPO Score	SLV Score	Citizen Satisfaction Score
PS Score	1			
MPO Score	0.183***	1		
SLV Score	0.0310***	0.0305***	1	
Citizen Satisfaction Score	0.0495***	0.0462***	0.0125**	1

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.2 "Truth" Scores

To measure the level of "true" PS performance, we start by mapping the KPIs under the PS self-reported measures to the indicators under MPO inspections. It is important to note that some KPIs cannot be verified for veracity accurately due to differing frequency of data collection. For example, while the PSes self-report the cleaning frequency of roads and drains in their panchayat, the MPOs cannot ascertain the cleaning frequency during their inspection visit. The MPOs can only ascertain the state of cleanliness of roads and drains during their visit.

This mapping exercise gives us an estimate of the "truthable" PS score; the proportion of the PS score can be accurately verified as true PS performance by the MPOs. The remaining portion of the self-reported PS score can be attributed as inflation on the part of PSes, to misreport their performance. Hence, we define score inflation as the difference between self-reported PS score and the score awarded to the same PS by the MPO, based on their inspections.

Missing data on KPIs, either from the PS or from the MPO, poses a challenge in our ability to validate PS performance. To account for this, we define an "adjusted PS score" on the

basis of whether both the PS and MPO have reported a score for each KPI. A KPI is said to be "non-truthable" if the MPO has not entered a score for it, during the inspection. So the respective portion of the PS score for that indicator is subtracted from the sum total of the PS score, to obtain the adjusted PS score. Out of a total PS score worth 100 points we determined 52 points were theoretically "truthable". Missing data further reduced the average "truthable" score to 34. For the SLV data the form was even more limited so there were only 18 points theoretically "truthable." Missing data further reduced the average truthable score is 13.97.

4.3 Score Inflation

Since the self-reported performance score is prone to inflation, we define and use three complementary measures of score inflation to monitor actual PS performance on the ground.

Total inflation is the percentage-point difference between the PS score percentage (PS points as a share of the total possible points on every KPI the PS reports in that period) and the MPO score percentage (MPO points as a share of the total possible points on every KPI rated by the MPO in that period).

$$\text{TotalInfl}\,(\%) = \left(\frac{\text{PS Score}}{\text{Max points on all PS KPIs}} - \frac{\text{MPO Score}}{\text{Max points on all MPO KPIs}}\right) \times 100$$

Verified inflation is the difference between the adjusted PS score and the adjusted MPO score computed only on KPIs observed in both sources during the same period, and is scaled by the total "truthable" points, where the truthable points are calculated as the sum total of all KPIs for which there are non-missing observations in the PS and MPO data.

$$\text{VerifiedInfl}\,(\%) = \left(\frac{\text{Adjusted PS score} - \text{Adjusted MPO score}}{\text{Total truthable points}}\right) \times 100$$

Finally, the *unverified inflation* proxies the inflation in the PS score that comes from KPIs without a counterpart in the MPO score. It equals the PS score on these unmatched KPIs expressed as a percentage of the total points available for those KPIs, minus the overall MPO score expressed on the same percentage-point scale.

$$\text{UnverifiedInfl}\,(\%) = \left(\frac{\text{PS points on PS-only KPIs}}{\text{Max points on PS-only KPIs}} \times 100\right) - \text{MPO score \%}$$

Due to the carefully constructed administrative data sources discussed above we are able to measure score inflation directly. We find that there is substantial inflation in our setting. In GPs that didn't receive the scorecard, the median total inflation was 15 percentage points.

4.4 Effort Metric

A unique feature of the mobile application is the requirement that panchayat secretaries submit a daily sanitation report which the app would only allow to submit if the phone's GPS was within the village. While, in practice, some PSes were able to find subordinates or other technical workarounds, reporting is a good proxy for PS effort. Daily physical presence at the GP is a crucial and costly part of the job since nearly all PSes do not reside in the villages in which they serve.

5.1 Balance

The treatment and control groups are generally balanced on observables. Importantly, the experiment is balanced with respect to the outcome variables in the April "baseline" period. Note that due to government hiring rules there are a substantial amount of PSes that are woman or come from an SC/ST background. These PSes from marginalized backgrounds may face further challenges dealing with village elected leaders.

Variable	(1) Control N/[Clusters]	Mean/SE	(2) Treatment N/[Clusters]	Mean/SE	T-test Difference (1)-(2)
No. of Female PSes	18326	0.316	55461	0.289	0.028**
	[134]	(0.012)	[400]	(0.007)	
No. of Female MPOs	18326	0.165	55461	0.126	0.039
	[134]	(0.033)	[400]	(0.018)	
JPS	18326	0.796	55461	0.798	-0.002
	[134]	(0.010)	[400]	(0.006)	
No. of regular PSes	18326	0.984	55461	0.986	-0.001
	[134]	(0.003)	[400]	(0.002)	
PS experience (in days)	18228	883.297	55335	912.948	-29.651
	[134]	(19.947)	[400]	(13.505)	
MPO experience (in days)	18326	838.020	55356	899.428	-61.408
	[134]	(59.423)	[399]	(48.907)	
No. of regular MPOs	18326	0.901	55461	0.933	-0.032
	[134]	(0.030)	[400]	(0.014)	
Mandal size	18326	27.135	55461	26.807	0.329
	[134]	(0.902)	[400]	(0.485)	
No. of MPW					
workers per village	18221	4.595	55167	4.361	0.234
	[134]	(0.184)	[400]	(0.089)	
Average no. of roads					
per village	18326	18.389	55461	17.544	0.845
	[134]	(0.800)	[400]	(0.468)	
Average no. of drains					
per village	18326	9.547	55461	8.769	0.778*
	[134]	(0.521)	[400]	(0.295)	
MPO Score	1317	54.842	3957	55.858	-1.016*
	[133]	(0.640)	[395]	(0.397)	
PS Score	2613	67.125	7915	66.954	0.171
	[134]	(0.414)	[400]	(0.264)	
Score inflation	1317	0.621	3955	0.259	0.362
	[133]	(0.370)	[395]	(0.220)	
Score inflation					
Share of big inflators	18326	0.060	55461	0.053	0.007*
	[134]	(0.006)	[400]	(0.003)	
SLV score	2614	6.941	7916	6.819	0.122
	[134]	(0.660)	[400]	(0.353)	
SLV score inflation	241	11.297	732	14.922	-3.626**
	[89]	(2.127)	[266]	(1.094)	
Citizen satisfaction score	955	3.709	2870	3.766	-0.057***
	[134]	(0.020)	[400]	(0.013)	

Notes: Dependent variables are measured in April 2021. The value displayed for t-tests are the differences in the means across the groups. Standard errors are clustered at variable mandal. Fixed effects using variable district_id are included in all estimation regressions.

5 Main Results

Our primary specification is a basic difference-in-difference style estimator with standard errors clustered at the mandal level. We control for various covariates: PS gender, years of work experience as PS, PS caste, posting type of PS, MPO gender, years of experience as MPO, and posting type of MPO.

The difference in difference style estimator pools all of the treatment months together. For robustness we also run the event study style estimator. We find there are no significant month by month differences in treatment effects, but they do qualitatively appear to slightly attenuate over time (appendix)

We observe a GP i in month t in strata s. The treatment months June,July, August, and September are represented by T. The preferred specification is:

$$Y_{it} = \alpha + \beta_1 Treat * \mathbb{1}_{t \in T} + \beta_2 \mathbb{1}_{t \in T} + + \beta_3 X + \mu_t + \mu_m$$

where μ_t are month fixed effects and μ_m are mandal fixed effects.

We find significant but small increases in panchayat secretary scores. Note that while on a scale out of 100 an increase of less than 1 may seem insignificant the scores are highly clustered. The inter-quartile range is just 13 from 67 to 80 and the standard deviation is 10. However, the scorecards effect seems to be contained to the PSes themselves. The other bureaucrats inspecting GPs, MPOs and SLVs, do not report any differences in score between treatment and control GPs. In fact, the coefficients are, if anything, slightly negative. The negative effect on official scores matches the scorecards negative impact on citizen satisfaction. The effect translates to a 4% decrease relative to the control mean.

The logical explanation for the divergence in treatment effects between what the PSes report and the citizen assessment is that the treatment induces PSes to inflate their scores.

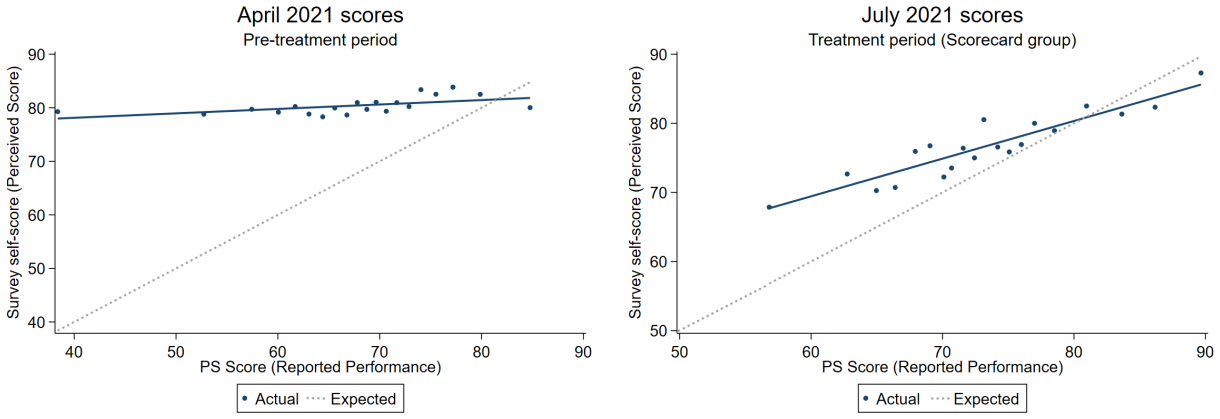
Table 3: Main Effects

Panel A:	(1) Self-Reported Assessment	(2) Supervisor Assessment	(3) Auditor Assessment	(4) Citizen Assessment
Score-cards x Post	1.019***	-0.377	-0.539	-0.145**
	(0.392)	(0.756)	(0.840)	(0.074)
Observations	255,300	111,253	17,363	11,138
Clusters	540	540	539	539
Mandal FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Panel B:	(1) Total Inflation	(2) Verified Inflation	(3) Unverified Inflation	(4) Effort Metric
Score-cards x Post	1.707**	1.387	1.328	-0.206
	(0.707)	(0.875)	(1.425)	(0.439)
Observations	89,330	89,330	73,622	1,059,910
Clusters	540	540	540	540
Mandal FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes

Column (1) are the PS scores from the app which theoretically range from 0 to 100, but practically have an inter-quartile range of 66 to 80. Supervisor assessments are the scores from the MPO app on a similar scale. Auditor assessments are scores from SLV's. Citizen assessments are scores from citizen phone surveys that simply asked for satisfaction on a 1-5 scale. The inflation measures are discussed in detail in section 4.3. The effort metric is a measure of whether the PSes attended their GPs every day.

6 Discussion

There is evidence that the scorecard intervention did help PSes learn more about their relative performance. In the April PS survey, PSes self-perceived performance bore no relationship to the PS scores. After the scorecards, PSes surveyed in July in the scorecard groups had scores that more closely matched their perceptions. Thus the scorecards did help some PSes better internalize there relative performance in the eyes of the department.



score inflation in general does not appear to be strongly correlated with observables such as gender, caste, and type of posting. Regressions examining the correlates of inflation can be found in Appendix F.

Misperception is one of the only variables that is has statistically significant heterogeneous treatment effects. These results suggest that PSes that stood to learn the most from scorecards were more likely to respond by actually performing better rather than inflating their scores.

Panel A:	(1) Self-Reported Assessment	(2) Supervisor Assessment	(3) Auditor Assessment	(4) Citizen Assessment
Score-cards x Post x Misperception	1.194**	2.447**	-0.479	0.355
	(0.576)	(1.021)	(1.925)	(0.255)
Post x Misperception	2.724***	-1.542*	0.813	-0.293
	(0.445)	(0.871)	(1.534)	(0.210)
Post x Score-cards	0.344	-1.556	0.344	-0.340**
	(0.430)	(1.008)	(1.556)	(0.170)
Observations	55,000	26,476	3,353	3,752
Clusters	538	538	430	482
Panchayat FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Panel B:	(1) Total Inflation	(2) Verified Inflation	(3) Unverified Inflation	(4) Effort Metric
Score-cards x Post x Misperception	-1.339	-0.431	-3.270	1.254*
	(1.127)	(1.203)	(2.778)	(0.680)
Post x Misperception	4.237***	3.255***	6.431***	1.019**
	(0.937)	(0.896)	(2.275)	(0.515)
Post x Score-cards	2.315**	0.984	2.741	-0.811
	(0.957)	(1.104)	(2.177)	(0.507)
Observations	21,200	21,200	17,334	228,250
Clusters	538	538	537	538
Panchayat FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes

Figure 2: Generic ML heterogeneity

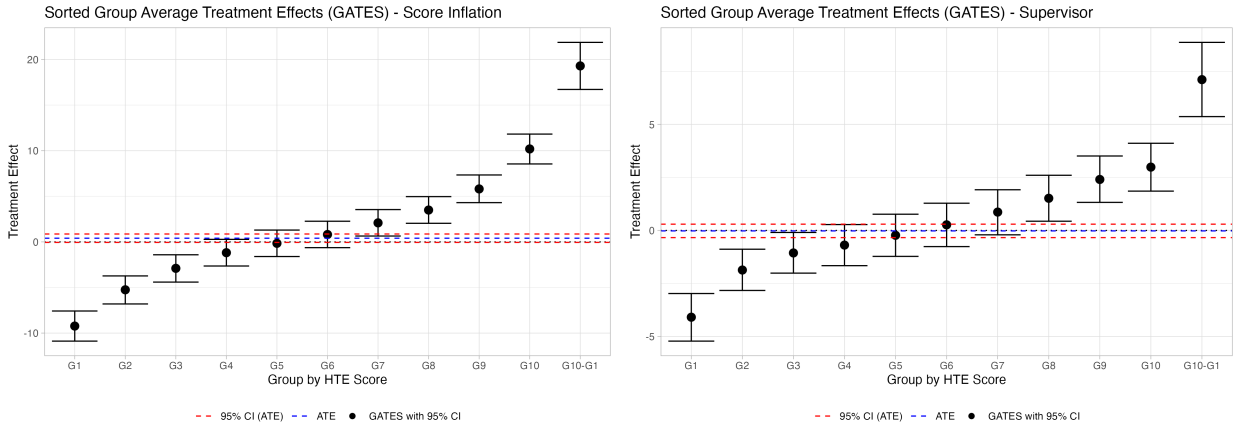


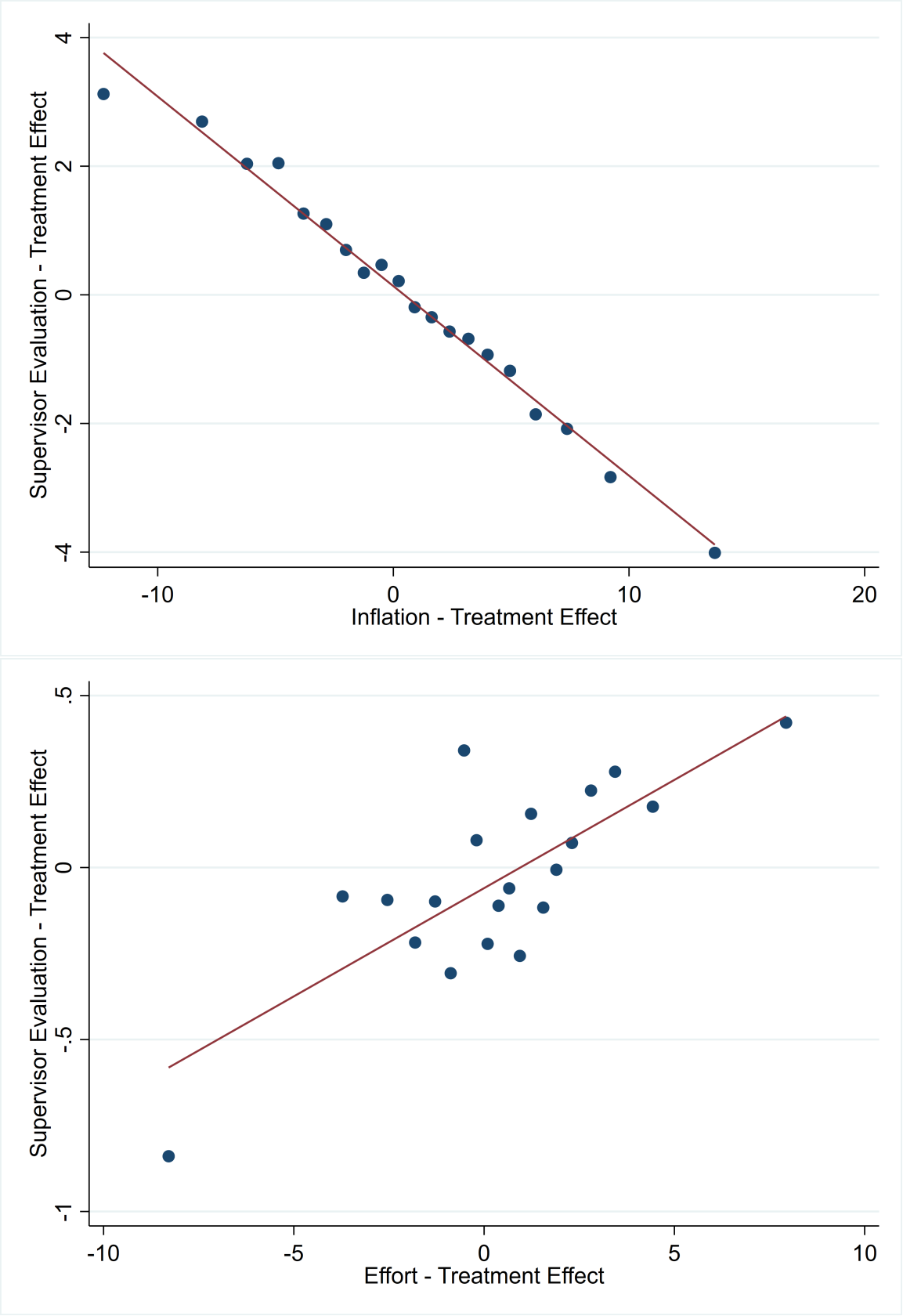
Table 4: Predicted treatment effects

Category	Performance	Data Accuracy	Share
Clearly positive	+	+	18%
Weakly positive	+	0	7%
Weakly positive	0	+	9%
Positive impact			34%
Tradeoff	+	-	4%
Tradeoff	-	+	2%
Weakly negative	-	0	13%
Weakly negative	0	-	12%
Clearly negative	-	-	15%
Negative impact			40%
No impact	0	0	19%

While demographic variables might not strongly predict heterogeneous effects on inflation individually, that does not mean there is not significant heterogeneity in treatment effects across PSes. In praticular, there may be complicated combinations of PS and MPO characteristics and baseline scores that are highly predictive. To test for this we followed the generic machine learning method proposed in (Chernozhukov, Demirer, Duflo, and FernÃ¡ndez-Val, 2017). The details of the method are discussed in appendix D. The basic finding was that the ML algorithm can sort PSes into groups that both significantly increased and decreased their scores in response to the scorecards treatment.

Notably, none of the variables used to generate individual treatment effects are good predictors of being in a high or low treatment group by themselves when comparing the standardized difference between the highest and lowest treatment effect groups for each of the predictors used (A).

Figure 3: Relationship between predicted treatment effects



The correlation of individual treatment effects across outcomes, but limited direct relationship with observables, highlights the challenge of performance management in a bureaucracy. Employees can differ dramatically in earnestness and intrinsic motivation, resulting in heterogeneous responses to performance management, but it is difficult for policymakers to classify employees into motivation types from observables alone.

7 Conclusion

Do bureaucrats use self-reported scores as an opportunity to make themselves look better in the eyes of their departments or as a tool to learn about their own performance? Our study suggests bureaucrats use scoring systems in both ways. While some bureaucrats were content to simply learn about their performance, others, especially those who learn their performance is relatively low, took the next step of "correcting" their scores in the following month. The fact that not all bureaucrats inflated their scores to the same degree makes it more difficult to confidently target resources and track progress using the self-reported data. One promising solution is to incentivize accuracy in addition to performance. Rather than simply rewarding bureaucrats for having the highest scores, governments should also reward bureaucrats whose data stands up to the scrutiny when inspected. If bureaucrats can inflate their scores in response to simple information about relative performance, then they can also report honestly in response to simple information about their relative accuracy. Policy solutions to encourage accurate self-reporting are important. Governments will struggle to deliver services at scale until they can trust and use data generated from the bureaucrats implementing their programs.

References

ACEMOGLU, D., L. FERGUSSON, J. ROBINSON, D. ROMERO, AND J. F. VARGAS (2020): "The perils of high-powered incentives: evidence from Colombia's false positives," American Economic Journal: Economic Policy, 12(3), 1–43.

BESLEY, T., R. BURGESS, A. KHAN, AND G. XU (2022): "Bureaucracy and development," Annual Review of Economics, 14, 397–424.

CHEN, Y. J., P. LI, AND Y. LU (2018): "Career concerns and multitasking local bureaucrats: Evidence of a target-based performance evaluation system in China," Journal of Development Economics, 133, 84–101.

CHERNOZHUKOV, V., M. DEMIRER, E. DUFLO, AND I. FERNÃ¡NDEZ-VAL (2017): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," .

DE JANVRY, A., G. HE, E. SADOULET, S. WANG, AND Q. ZHANG (2020): "Performance Evaluation, Influence Activities, and Bureaucratic Work Behavior: Evidence from China," .

DUFLO, E., R. HANNA, AND S. P. RYAN (2012): "Incentives work: Getting teachers to come to school," American Economic Review, 102(4), 1241–78.

FINAN, F., B. A. OLKEN, AND R. PANDE (2017): "The personnel economics of the developing state," Handbook of economic field experiments, 2, 467–514.

FISMAN, R., AND Y. 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.

GINÉ, X., G. MANSURI, AND S. A. SHRESTHA (2022): "Mission and the bottom line: Performance incentives in a multigoal organization," Review of Economics and Statistics, 104(4), 748–763.

JACOB, B. A., AND S. D. LEVITT (2003): "Rotten apples: An investigation of the prevalence and predictors of teacher cheating," The Quarterly Journal of Economics, 118(3), 843–877.

KHAN, A. Q., A. I. KHWAJA, AND B. A. OLKEN (2016): "Tax farming redux: Experimental evidence on performance pay for tax collectors," The Quarterly Journal of Economics, 131(1), 219–271.

KHAN, A. Q., A. I. KHWAJA, AND B. A. OLKEN (2019): "Making moves matter: Experimental evidence on incentivizing bureaucrats through performance-based postings," American Economic Review, 109(1), 237–70.

MURALIDHARAN, K., P. NIEHAUS, S. SUKHTANKAR, AND J. WEAVER (2021): "Improving last-mile service delivery using phone-based monitoring," American Economic Journal: Applied Economics, 13(2), 52–82.

OLKEN, B. A. (2007): "Monitoring corruption: evidence from a field experiment in Indonesia," Journal of political Economy, 115(2), 200–249.

RAO, V., K. ANANTHPUR, AND K. MALIK (2017): "The anatomy of failure: An ethnography of a randomized trial to deepen democracy in rural India," World development, 99, 481–497.

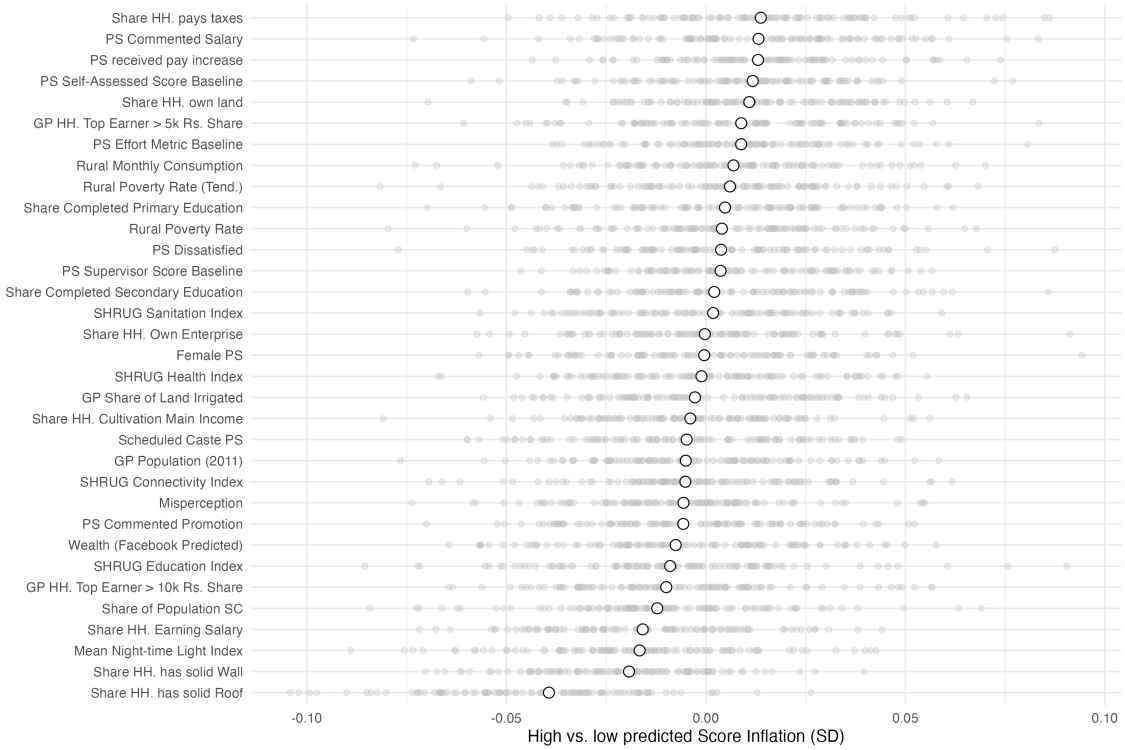
RASUL, I., AND D. ROGGER (2013): "Management of bureaucrats and public service delivery: Evidence from the Nigerian civil service," .

SINGH, A. (2020): "Myths of official measurement: Auditing and improving administrative data in developing countries," Research on Improving Systems of Education (RISE) Working Paper, 42.

VEERARAGHAVAN, R. (2021): Patching development: Information politics and social change in India. Oxford University Press.

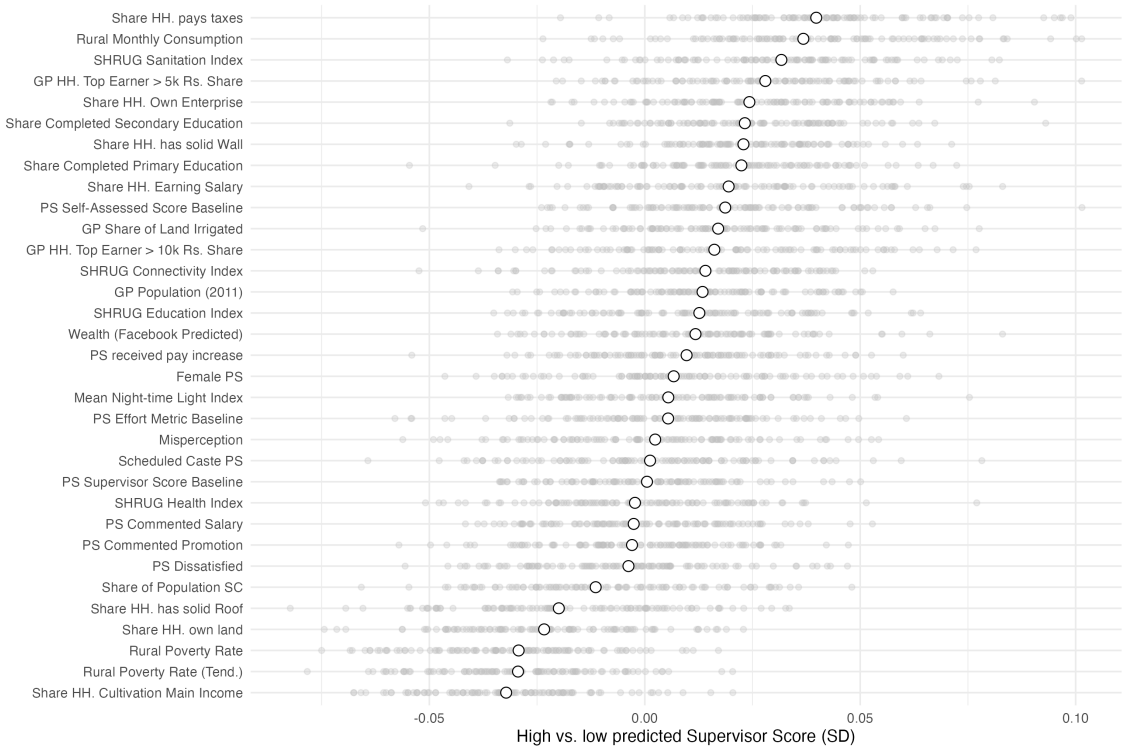
A Covariate Predictors of Individual Treatment Effects

Figure 4: Predictors of Low vs High Individual Inflation Treatment Effect



The graph plots the standardized difference between PSes in the highest and lowest quintile of predicted treatment effects for each baseline characteristic used in the random forest. Each dot plots one split for a given covariate. Large circles with white centers plot the median. Colored covariates have at least 90% or more splits above (or below) zero; gray covariates do not

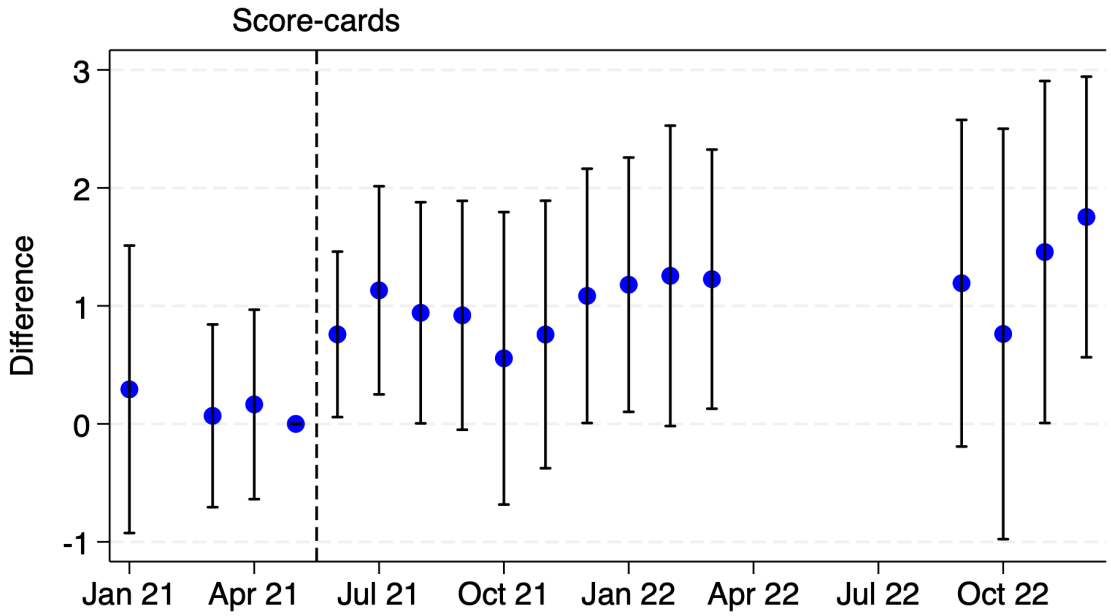
Figure 5: Predictors of Low vs High Supervisor Score Treatment Effect



The graph plots the standardized difference between PSes in the highest and lowest quintile of predicted treatment effects for each baseline characteristic used in the random forest. Each dot plots one split for a given covariate. Large circles with white centers plot the median. Colored covariates have at least 90% or more splits above (or below) zero; gray covariates do not

B Treatment Effect by Month

Figure 6: Effect of treatment on PS Score - By month



C PS Score Rubric

KRA 1: Sanitation (40 points)

1. Road cleaning frequency (4 pts): Proportional to the number of roads cleaned each day. The score will be calculated only taking into consideration the 18 best days in the month.

2. Road cleaning (7 pts): Proportional to the percentage of randomly selected (DSR) roads which are clean. The score will be calculated only taking into consideration the 18 best days in the month.

3. Drain cleaning frequency (4 pts): Proportional to the number of drains cleaned each day. The score will be calculated only taking into consideration the 18 best days in the month.

4. Drain cleaning (7 pts): Proportional to the percentage of randomly selected (DSR) drains which are clean.The score will be calculated only taking into consideration the 18 best days in the month.

5. Institution cleanliness (6 pts): Proportional to the percentage of randomly selected (DSR) institutions which are clean. The score will be calculated only taking into consideration the 18 best days in the month.

6. Garbage transport (4 pts): Proportional to the days in the month when garbage is transported to the segregation shed/ dumpyard. The score will be calculated only taking into consideration the 18 best days in the month.

7. Household waste segregation (4 pts): Proportional to the number of days in the month when segregated waste is collected from HHs (measured as days when segregated waste was collected from at least one HH)

8. Compost preparation (4 pts): 2 points if compost facilities present; 2 points if compost is prepared (from MAS)

KRA 2: Streetlights (8 points)

1. Streetlights (8 pts): Streetlights scores are based on the share of streetlights which are functional on an average day in the month (in the best 18 days). If this average is lower than 0.6, the score is 0, and increases linearly beyond that.

KRA 4: Greenery (20 points)

1. Nursery infrastructure (6 pts): 2 points each if the village nursery has watering facility, cattle trap, fencing.
2. Plantation watering (5 pts): Equal points (5/3) given if each of the 3 plantation types are watered regularly.
3. Plantation fencing (2 pts): Equal points (2/3) given if each of the 3 plantation types have fencing
4. Plantation survival (7 pts): Proportional to plantation survival rates, equally divided for the 3 plantation types.

KRA 5: Administration (12 points)

1. GP/GS meetings (3 pts): 2 points awarded if GP meeting held in the past month, 1 point awarded if GS meeting held in the past 2 months.
2. Register maintenance (2 pts): Proportional to the number of registers which are up to date.
3. Birth and death registration and certification (3 pts):

   0.75 points each are proportional to the number of birth and death registered within timelines

   0.75 points each are proportional to the number of birth and death certificates issued within timelines
4. E-panchayat computerization (2 pts): Proportional to the number of days in the month that the DSR has been filled.
5. CC charges paid on time (2 pts): Proportional to the share of bill amount for which cheque has been generated.

KRA 6: NREGA (18 points)

1. NREGA (18 pts): Proportional to the days of work per reported wage seeker. Graded relative to the 90th percentile GP in the mandal (GPs at and above the 90th percentile get full marks). GPs with less than or equal to 5 reported wage seekers are not eligible for NREGA score, and their overall PS score will be graded out of 82 instead of 100.

D Generic ML

We estimate heterogeneous treatment effects using generic machine learning, a method proposed in (Chernozhukov, Demirer, Duflo, and FernÃ¡ndez-Val, 2017). This is a tool-independent procedure to facilitate the inference of individual-specific heterogeneous treatment effects in RCT settings. In this section, we briefly describe the generic ML procedure and highlight the results that emerge from applying this method in our setting.

The main stages of GenericML are: (i) estimate HTE via machine learning models with train/test data split, (ii) compute linear homogeneous TE proxies (BLP/CATE) from the machine learning models, (iii) rank observations by the predicted effect size (GATES); (ic) test difference in baseline characteristics (CLAN) between groups with the highest and lowest effect sizes.

GenericML makes the HTE estimation possible by predicting the counterfactual outcomes with ML models. It does so by repeating the following steps with a set of candidate ML models:

1. randomly 50/50 split the full data set into a main set M and an auxiliary set A. A and M both contain treated and untreated subject

2. train ML models with only samples from M to formulate the relationship between baseline characteristics and the outcome. In particular, we want learn such relationship respectively in the control and the treatment group.

3. Having formulated the relationships, we then can 'see' the counterfactual outcome for each individual in A (who was not involved in training models). Simply, for each individual a in A, we can predict its: treated outcome (by plugging its baseline characteristics in f_t, denote the predicted outcome as $f_t(a)$), control group outcome (by plugging in f_c, denote as $f_c(a)$) and predicted treatment effect (by taking the difference, HTE = $f_t(a) - f_c(a)$).

After repeating the above steps with multiple random A/M splitting each time, we have the statistics of HTE derived by apply different ML model, and move to the next step - horse racing between candidate models.

4. We obtain HTE estimates for each observation by pretending it is either in control and treatment group. However, each observation is actually either in the treatment or the control group, and its observed outcome is known. We can therefore assess prediction quality of each model by comparing the observed to the predicted

outcome. Averaging such comparison across different individuals lead to a performance indicator for a model. We then proceed with the model of the highest performance indicator.

5. GATES: having fixed the best model, we predict HTE with that model and sort the individuals according to predicted size of treatment effect. We then split the individuals into 5 groups ranked by effect size; the average treatment effect is then estimated for each group.

GenericML has two main advantages over simpler methods of estimating heterogeneous treatment effects, such as including interactions between treatment dummies and particular covariates. First, GenericML captures more complex heterogeneity that sequential addition of interaction terms

By horse-racing multiple prediction models and including non-parametric models, GenericML enables data-driven model choice. Instead of discretionary choice of model, GenericML provides a data-driven approach to automated model selection. Moreover, GenericML pipeline makes the horse-racing of models feasible and, more importantly, credible by enforcing a shared performance indicator, based on which we decide which model is the best. With inclusion of non-parametric ML models, GenericML features better robustness in HTE estimation against high-dimensional data where the true data generating process is indeed high-dimensional. Consider the old-school way of detecting heterogneity: by introducing interaction terms in regression analysis. When there are many regressors, fully saturated regressions are likely to be overfit and absorb too much noise.

E Correlates of Supervisor Score

There is some evidence that supervisors slightly discriminate against female panchayat secretaries within their mandals.Lower MPO scores for SC PSes seem to be driven by where the SC PSes serve as significant differences disappear when strata, which include district, or mandal fixed effects are added to the model.

F Correlates of Inflation

	(1) MPO Score	(2) MPO Score	(3) MPO Score	(4) MPO Score
PS Gender	0.145	0.136	-0.0226	0.323***
	(0.169)	(0.167)	(0.872)	(0.112)
JPS	0.0320	0.00458	0.942	-0.0172
	(0.191)	(0.191)	(0.865)	(0.116)
PS Caste	-0.148***	-0.144**	-0.129	-0.0312
	(0.0564)	(0.0563)	(0.249)	(0.0337)
Observations	31971	31971	15481	31971
Mean of depvar	28.02	28.02	28.50	28.02
Month FE	No	Yes	Yes	Yes
Strata FE	No	No	Yes	No
Mandal FE	No	No	No	Yes

Standard errors in parentheses

SEs are clustered at mandal-level.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Effect of PS' Caste on score inflation percentage

	(1) Score inflation %	(2) Score inflation %	(3) Score inflation %	(4) Inflation >25pc
Minorities	0.307	0.328	0.160	-0.00623
	(1.790)	(1.447)	(1.434)	(0.0193)
OC	-0.224	0.285	0.422	-0.000291
	(0.592)	(0.504)	(0.501)	(0.00863)
SC	0.501	0.0734	0.0468	0.00571
	(0.488)	(0.407)	(0.404)	(0.00656)
ST	-0.0635	-0.611	-0.535	-0.00586
	(0.684)	(0.578)	(0.583)	(0.00865)
Observations	24413	24413	24413	24413
Mean of depvar	3.950	3.950	3.950	0.131
Strata FE	No	Yes	No	Yes
Month FE	No	No	Yes	Yes

Standard errors in parentheses

SEs are clustered at mandal level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Effect of PS' gender on score inflation percentage

	(1) Score inflation %	(2) Score inflation %	(3) Score inflation %	(4) Inflation >25pc
PS Gender	-0.522	-0.552	-0.554	0.00148
	(0.415)	(0.377)	(0.377)	(0.00613)
Observations	24413	24413	24413	24413
Mean of depvar	3.950	3.950	3.950	0.131
Strata FE	No	Yes	Yes	Yes
Month FE	No	No	Yes	Yes

Standard errors in parentheses. Female coded as 1

SEs are clustered at mandal level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$