American Economic Journal: Applied Economics 2025, 17(3): 414–445
https://doi.org/10.1257/app.20230672

Information Systems, Service Delivery, and Corruption: Evidence From the Bangladesh Civil Service†

By Martin Mattsson*

Slow public service delivery and corruption are common problems in low- and middle-income countries. Can better management information systems improve delivery speed? Does improving the delivery speed reduce corruption? In a large-scale experiment with the Bangladesh Civil Service, I send monthly scorecards measuring delays in service delivery to government officials and their supervisors. The scorecards increase on-time service delivery by 11 percent but do not reduce bribes. Instead, the scorecards increase bribes for high-performing bureaucrats. A model where bureaucrats' reputational concerns constrain bribes can explain the results. When positive performance feedback improves bureaucrats' reputations, the constraint is relaxed, and bribes increase. (*JEL* D73, D83, H83, O17)

A government's capacity to implement its policies, secure property rights, and provide basic public services is paramount for economic development. To have this capacity, states need functioning bureaucracies with government officials motivated to carry out these tasks. While explicit incentive structures such as pay-for-performance contracts can change the behavior of government officials, they are often hard to implement without unintended consequences or political resistance (Banerjee, Duflo, and Glennerster 2008; Dhaliwal and Hanna 2017). Another approach is to improve systems that measure bureaucrats' performance. This may improve incentives by allowing supervisors to let job performance determine postings and promotions, a strong motivator for civil servants (Khan, Khwaja, and Olken 2019; Bertrand et al. 2020; Deserranno, Kastrau, and León Ciliotta 2022). Regular performance feedback may also make the performance more salient to the bureaucrats themselves, potentially leveraging government officials' intrinsic motivation (Prendergast 2007; Banuri and Keefer 2016).

* Mattsson: National University of Singapore (email: martin.mattsson@nus.edu.sg). Benjamin Olken was the coeditor for this article. I would like to thank Nathan Barker, Gaurav Chiplunkar, Anir Chowdhury, Andrew Foster, Eduardo Fraga, Sahana Ghosh, Kenneth Gillingham, Marina Halac, Johannes Haushofer, Ashraful Haque, Enamul Haque, Daniel Keniston, Mushtaq Khan, Ro'ee Levy, Edmund Malesky, Imran Matin, Mushfiq Mobarak, Rohini Pande, Dina Pomeranz, Mark Rosenzweig, Nick Ryan, Yogita Shamdasani, Jeff Weaver, Jaya Wen, Julian Wright, Fabrizio Zilibotti, and three anonymous referees as well as numerous seminar participants for helpful comments and suggestions. I also thank Saadman Chowdhury, Muhammad Bin Khalid, Mahzabin Khan, and Ashraf Mian for excellent research assistance and IPA Bangladesh for outstanding research support. A randomized controlled trial registry and preanalysis plan are available at www.socialscienceregistry.org/trials/3232 (Mattsson 2021). This project was approved by Yale University IRB (Protocol ID 2000021565). This work was supported by the JPAL Governance Initiative (GR-0861), the International Growth Centre (31422), the Yale Economic Growth Center, the MacMillan Center for International and Area Studies, the Weiss Family Fund, the Sylff fellowship, and National University of Singapore Start-Up Grant.

† Go to https://doi.org/10.1257/app.20230672 to visit the article page for additional materials and author disclosure statement(s).

This paper studies an information system designed to improve the processing times of applications for changes to government land records in Bangladesh. Both slow public-service delivery and corruption are substantial problems in Bangladesh. Among households that paid a bribe for a public service, 23 percent stated that "timely service" was one of the reasons for the bribe (Transparency International Bangladesh 2018). This suggests that policies and technologies that speed up service delivery on average may also reduce corruption, as fewer citizens and firms would need to pay bribes to receive the services within the time in which they need them. However, such an effect has not been established empirically.

In an experiment with the Bangladesh Civil Service, I provide information on junior bureaucrats' performance using scorecards sent to the bureaucrats and their supervisors every month over a 16-month period. The scorecards are designed to reduce delays in processing applications for land-record changes and are based on data from an e-governance system. Two performance indicators appear on the scorecards: the number of applications processed within a time limit of 45 working days and the number of applications pending beyond that limit. The scorecards also show bureaucrats' performance on these indicators relative to all other bureaucrats in the experiment. Each bureaucrat manages a subdistrict land office and the intervention is randomized at the office level. The experiment covers 311 land offices (60 percent of all land offices in Bangladesh), serving a population of approximately 97 million people.

The scorecards improve processing times. Using administrative data from more than a million applications, I estimate that the scorecards increase the share of applications processed within the time limit by 6 percentage points (11 percent) and decrease processing times by 13 percent. The effect is present throughout the 16 months of the experiment and is driven by improvements among offices that were underperforming relative to the median at baseline. Among the bureaucrats that were underperforming at the start of the experiment, I also find a 17 percentage point negative effect on promotions 12–18 months after the experiment. This suggests that the scorecards allowed supervisors to better align the frequency of promotions with performance, thus improving the incentives for bureaucrats.

Despite their effect on processing times, the scorecards did not decrease bribe payments. I collect survey data on bribe payments from applicants; the point estimate for the effect on my measure of bribes paid is an increase of BDT 940 (US$ 11 or 15 percent). The lower bound of the 95 percent confidence interval is a decrease of 4 percent; thus, the result is inconsistent with a substantial reduction in bribes paid. Using a randomized information intervention among surveyed applicants, I rule out that lack of information about the improved processing times is the reason for not seeing a negative effect on bribes paid. This suggests that interventions targeting processing times, even if successful, may not reduce corruption.

The positive effect of the scorecards on bribe payments is concentrated among the offices that were overperforming relative to the median at baseline, where the scorecards did not affect processing times. In the underperforming offices, where scorecards improved processing times, there is no effect on bribes.

To explain these results I propose a model in which bureaucrats trade off reputation, bribe money, and the utility cost of effort. Their reputation is determined by

their visible job performance along two dimensions, processing times and bribes taken. The scorecards are modeled as an increase in the visibility of processing times, making processing times more important for reputation, which in turn incentivizes all bureaucrats to improve their processing times (akin to a substitution effect).[1] However, for overperforming bureaucrats, the increased visibility of their good performance also increases their reputation level, which reduces the marginal importance of reputation (akin to an income effect). For overperforming bureaucrats, the two effects run in opposite directions, and the overall effect predicted by the model on processing times is ambiguous. For underperforming bureaucrats, the two effects run in the same direction, and the model predicts improved processing times.

In the model, bureaucrats refrain from taking more bribes because bribes negatively affect their reputation. When the scorecards improve the reputation level of overperforming bureaucrats, the marginal importance of reputation for their utility decreases, and this causes them to take more bribes. For underperforming bureaucrats, the decrease in reputation from their poor performance being more visible is counteracted by their increased effort. Therefore, the effect on bribes is ambiguous for underperforming bureaucrats. I also discuss several alternative explanations for the results such as improved service delivery increasing the willingness to pay by applicants, the scorecards affecting bureaucrats' perceptions of government priorities and the rate at which bureaucrats are transferred, and bureaucrats using the scorecards in negotiations over bribe payments with the applicants. While I cannot completely rule out all of these explanations, none of them can fully explain the empirical results.

This paper contributes to four strands of literature. First, it provides empirical evidence on the causal effect of a policy improving the speed of public-service delivery on corruption. It has been shown that slow public-service delivery is positively associated with corruption and that individuals seeking services may have to pay bribes to reduce the time between application and provision (Kaufmann and Wei 1999; Bertrand et al. 2007; Freund, Hallward-Driemeier, and Rijkers 2016). In the theoretical literature, one view is that corruption allows individuals to circumvent excessive bureaucratic hurdles (Leff 1964; Huntington 1968). An opposing view is that corruption causes delays and red tape in public services, as making the de jure regulation more onerous allows government officials to extract more bribes (Rose-Ackerman 1978; Kaufmann and Wei 1999; Mattsson 2023).[2] According to both views, we could reduce corruption by improving the speed of service delivery for everyone. However, I show that an intervention successfully targeting delays in service delivery did not decrease bribe payments in this context.

Second, this paper contributes to the literature on how incentives shape bureaucratic performance and corruption. There is an extensive literature on both monetary and nonmonetary explicit incentives (e.g., Duflo, Hanna, and Ryan 2012; Ashraf, Bandiera, and Jack 2014; Khan, Khwaja, and Olken 2016; Rasul and

[1] The increase in visibility can be interpreted as an improvement in the supervisors' ability to monitor this aspect of the bureaucrats' work, an increase in the salience of the information to the bureaucrats themselves, or both.

[2] In Banerjee (1997); Guriev (2004); and Banerjee, Hanna, and Mullainathan (2012), both corruption and red tape emerge from the nature of public service provision due to the principal agent problem between the government and its bureaucrats. The experimental results can neither reject nor provide evidence in favor of these models.

Rogger 2018; Khan, Khwaja, and Olken 2019; Khan 2023), including the role of reputation as an incentive (Leaver 2009). There is also a growing literature on the effects of information systems within government bureaucracies (Dhaliwal and Hanna 2017; Callen et al. 2020; Banerjee et al. 2021; Dal Bó et al. 2021; Dodge et al. 2021; Muralidharan et al. 2021; de Janvry et al. 2022; Raffler 2022; Debnath, Nilayamgode, and Sekhri 2023). Consistent with this literature, I show that increased transparency about individual civil servants' performance can improve public service delivery, even without explicit incentives, and that this effect is persistent over time. I provide evidence for a career concerns mechanism where bureaucrats avoid negative scorecards as those increase the time until their next promotion, but the effect could also be amplified by bureaucrats' sense of shame or pride.

Third, this paper contributes to the literature on the effects of performance monitoring. The finding that providing performance feedback leads to improvements in performance, especially for underperformers, is consistent with the results of several other experiments (e.g., Allcott 2011; Ayres, Raseman, and Shih 2013; Byrne, Nauze, and Martin 2018; Barrera-Osorio et al. 2020), including recent work with Indian bureaucrats by Dodge et al. (2021).[3] Most papers in this literature show smaller or even negative effects for high-performers. This paper adds to this literature by showing that these negative effects can spill over into domains not covered by the performance information, for which data are not typically collected. There is a substantial theoretical and empirical literature on the multitasking problem, that is, how incentives for improving one indicator have negative spillovers by taking attention and resources away from other types of performance (Holmström and Milgrom 1991; Finan, Olken, and Pande 2017). My model suggests a different mechanism for negative spillovers, namely decreasing the marginal utility from reputation after receiving positive performance feedback.[4]

Fourth, this paper contributes to the literature on the determinants of bribe amounts. Some models and empirical evidence suggest that bribe payers' outside options and abilities to pay constrain bribe amounts (Svensson 2003; Bai et al. 2019), potentially leaving little room for applicant complaints or government monitoring to reduce corruption (Niehaus and Sukhtankar 2013). In other settings, monitoring has been effective in reducing corruption (Reinikka and Svensson 2005; Olken 2007).[5] I show that, in my context, bribes are not fully determined by applicants' willingness to pay. My model highlights how bureaucrats' concerns for their reputations constrain bribes, thus explaining how bribes can be substantially below applicants' willingness to pay for the service.

[3] Ashraf, Bandiera, and Lee (2014) and Ashraf (2022) show that privately provided social comparisons reduced the performance of low-performing health care and garment workers, while publicly announcing good performances increases performance for low-performing groups. Blader, Gartenberg, and Prat (2020) show similar effects among truck drivers, but the results are reversed when the intervention is combined with a management practice establishing a cooperation-based value system.

[4] This is also consistent with the literature on moral licensing, showing that when past prosocial behavior is made more salient, individuals tend to act less altruistically (Sachdeva, Iliev, and Medin 2009; Clot, Grolleau, and Ibanez 2018).

[5] Do, Van Nguyen, and Tran (2021) show that doctors' fear of punishment, reputational concerns, or moral obligations cause them to take smaller bribes when treating patients with acute conditions.

The rest of this paper is organized as follows. Section I describes the context, experimental interventions, and data. Section II describes my empirical strategy. Section III presents the effects of the scorecards on processing times and bribes. Section IV discusses mechanisms and the effect on promotions. Section V describes the model of bureaucratic behavior and how it can explain the empirical results. Section VI concludes.

I. Context, Experimental Intervention, and Data

A. *Land-Record Changes in Bangladesh*

This paper studies land-record changes, called mutations in Bangladesh. When a parcel of land changes owners, either through sale or inheritance, the land record has to be updated and a new record of rights (*khatian*) issued to the new owner. An updated record of ownership is crucial for maintaining secure property rights. Unfortunately, the burdensome and costly process of applying for land-record changes is causing many new landowners to wait to apply until they need a record of rights. This means that the government land records substantially lag actual ownership, contributing to land disputes. Land disputes are the most severe legal problem in Bangladesh—where 29 percent of adults have faced a land dispute in the past four years (Hague Institute for Innovation of Law 2018).

Structure of the Bureaucracy.—Applications for land-record changes are processed by civil servants holding the position of assistant commissioner land (ACL). Throughout the paper, I refer to ACLs as *bureaucrats*. ACL is a junior position in the Bangladesh Administrative Service, the elite cadre of the Bangladesh Civil Service.[6] Each subdistrict (*upazila*) land office is headed by a single ACL, and processing land-record changes is a central duty of the ACL. In qualitative interviews, ACLs estimate spending between 25–50 percent of their working time on land-record changes. Bureaucrats typically hold an ACL position for one to two years.

As is common for most civil servants, the bureaucrats' pay is unrelated to their performance. Furthermore, it is extremely rare for bureaucrats to be suspended or dismissed. One of the few explicit incentives bureaucrats face are changes to the time between promotions and the attractiveness of future postings. See Supplemental Appendix B.1 for a discussion about the associations between performance, bribe taking, and bureaucrats' future career paths. After being promoted, bureaucrats will hold the rank of senior assistant commissioner, and typically their next posting will be unrelated to land administration.[7]

[6] In rare cases, more senior bureaucrats hold the ACL position. This happens in particularly important subdistricts or while waiting for the position to be filled by a junior bureaucrat.

[7] Assistant commissioners have a pay scale of BDT 22,000–53,060, while senior assistant commissioners have a pay scale of BDT 35,500–67,010. These basic salaries are supplemented with several allowances and benefits, which typically increase approximately proportionally to salary. An additional benefit of a promotion is that the bureaucrat will be in a better position for future promotions and more likely to reach the highest echelons of the civil service (Bertrand et al. 2020).

Although promotions in any given year follow a "tournament structure," it is not the case that only those close to the cutoff have a chance to get promoted (Deserranno, Kastrau, and León Ciliotta 2022). Instead, the effort that bureaucrats put toward getting promoted in a particular year is likely to carry over to a higher promotion probability in the next year and ultimately higher probability of reaching further in their career overall.

The ACL is directly supervised by a Upazila Nirbahi Officer (UNO), the most senior civil servant at the subdistrict level. During periods when no ACL is assigned to an office, the UNO is responsible for the ACL's duties. The UNO has substantial power to influence the ACL's future career as it is the UNO who writes the Annual Confidential Report about the ACL's performance. The UNO is in turn supervised by a deputy commissioner (DC), the most senior bureaucrat at the district level. Throughout the paper, I refer to the UNOs and DCs as *supervisors*.

Application Process.—The de jure process for making a land-record change is depicted in Supplemental Appendix Figure D1. The process starts when the new owner applies at the subdistrict land office. There is no competition between land offices for applicants, as each parcel of land is under the jurisdiction of a single subdistrict. The application is inspected by the office staff, who verify that the application has the required documents. The application is then sent to the local (Union Parishad) land office, which is the lowest tier of land offices. There, a land office assistant verifies the applicant's claim to the land by meeting with the applicant and visually inspecting the land. The Land Office Assistant then sends a recommendation back to the subdistrict land office on whether to accept or reject the application. The application is then verified against the government land record. Finally, the ACL holds a meeting with the applicant where the application is formally approved. The applicant then pays the official fee of 1,150 BDT (US$14) and receives the new record of rights.[8]

The government has mandated that applications should take no more than 45 working days to process, but in practice, delays beyond this time limit are common. In my data, only 56 percent of applications in the control group were processed within the time limit, and the average processing time was 64 working days.

Bureaucrats' Discretionary Powers and Corruption.—In practice, applicants also pay bribes. Transparency International Bangladesh (2016) estimates that the land sector is the second-largest receiver of bribes from Bangladeshi citizens and that the average bribe for a land-record change is 4,085 BDT. Supplemental Appendix Figure D2 shows that among the applicants in my survey, the average estimated bribe payment for "a person like themselves" is 6,718 BDT (US$80 and 1.5 months of the sample's average per capita household expenditure).

Supplemental Appendix Figure D3 shows that the most common responses to the open-ended question of why a bribe was paid are akin to: "to get the work

[8]Throughout the paper, I use a US$/BDT exchange rate of 84.3, the average exchange rate during the experiment.

done" (65 percent), "for faster processing" (12 percent), and "to avoid hassle and inconveniences" (6 percent). This highlights the bureaucrats' power over applications along two dimensions. First, they can decide whether to accept or reject the application. Second, they can speed up or slow down the application, as well as create various hassles for the applicant.

Supplemental Appendix Figure D2 shows that the average stated valuation of the land is 1,897,806 BDT (US$22,513). This is more than two orders of magnitude larger than the estimate of the average bribe payment. Applicants' stated willingness to pay for having their application processed within the shortest realistic processing time (seven days) is, on average, BDT 2,189 (US$26). Because this amount is lower than most of the estimates for the average bribe, it suggests that applicants are paying bribes not just for faster processing but also for getting the approval.

E-governance System for Land-Record Changes.—In February 2017, a new e-governance system for land-record changes was gradually introduced to simplify the process for both applicants and bureaucrats.[9] The e-governance system generates administrative data on each application made in the system. However, until the start of the experiment, these data were not used for evaluating bureaucrats.

B. *Experimental Intervention: Performance Scorecards*

The main experimental intervention consists of monthly scorecards designed to decrease delays in the processing of applications for land-record changes. I designed the scorecards in close collaboration with the Ministry of Land and a2i, the government agency responsible for the e-governance system. The intervention was randomized at the land-office level, but as there is only one bureaucrat (ACL) per land office, each scorecard is addressed directly to a bureaucrat.

Figure 1 depicts an example scorecard. The scorecard evaluates the bureaucrat's performance using two performance indicators: the number of applications processed within 45 working days in the past month, where a higher number indicates a better performance, and the number of applications pending beyond 45 working days at the end of the month, where a lower number indicates a better performance. The scorecards compare these numbers with the average numbers for all land offices in the experiment.[10] The scorecard also provides the office's percentile ranking for each indicator, with a short sentence and a thumbs-up or thumbs-down symbol reflecting the performance.

At the bottom of the scorecard, the number of applications received by the office over the past six months is displayed. Offices vary substantially in terms of

[9] The system was new at the time of the experiment, and not all applications were processed through it. In the survey data, 76 percent of applicants stated they made their application using the e-governance system. There is neither an overall effect of the scorecards on this share nor any heterogeneous effects by baseline performance. However, the share is 12 percent higher among offices overperforming at the start of the experiment (p-value 0.085). In qualitative interviews with bureaucrats, the most common reason for not using the e-governance system was that the relevant local (Union Parishad) land office had not yet installed the system.

[10] The initial comparison group was the 112 offices in the first randomization. After the second randomization, the group was expanded to include all 311 offices.

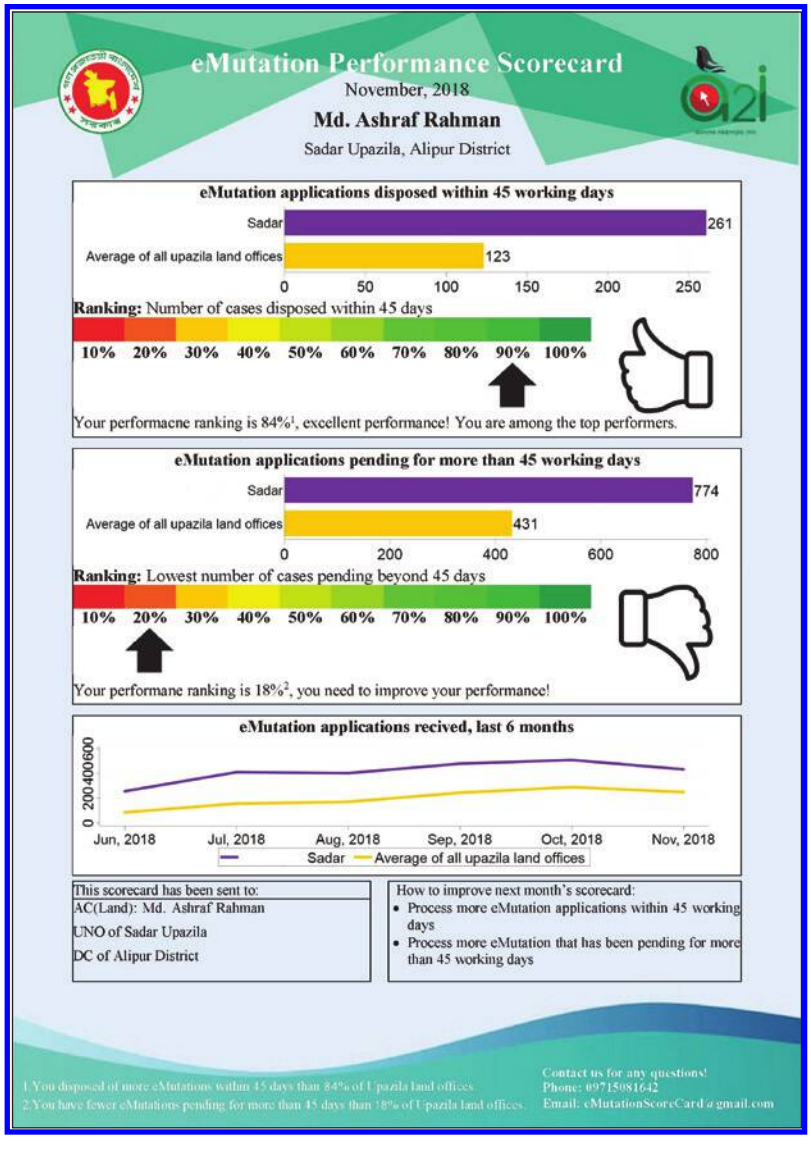

FIGURE 1. EXAMPLE OF A PERFORMANCE SCORECARD

Notes: Example of a performance scorecard in English. The scorecards were delivered every month during the 16 months of the experiment. The bureaucrat name and land office name are changed to preserve anonymity. The English scorecard was accompanied by a scorecard in Bengali as well as an explanatory note showing how the numbers are calculated. See discussion in Section IB and Supplemental Appendix B.4.

their workloads and the two performance indicators were chosen such that neither large nor small offices would have an inherent advantage in receiving a good average score, as shown in Supplemental Appendix Figure D4. Supplemental Appendix B.4 provides more details about the rationale for choosing the performance indicators and analyzes the extent to which bureaucrats matter for the performance score.

At the start of each month, I generated the scorecards using data from the e-governance system.[11] Hard copies of the scorecards in English and Bengali were then sent out by courier to all offices in the treatment group. An email with a PDF version of the scorecard was also sent to bureaucrats who had listed their email addresses publicly or in the e-governance system. Copies of the scorecards were also sent to the offices of the treated bureaucrats' supervisors (UNO and DC), which we mentioned in the scorecard itself to ensure that the bureaucrats were aware that their supervisors had received a copy.

Offices in the treatment group were not informed that they would receive a scorecard before the start of the treatment, but the first scorecard was followed by phone calls to the bureaucrats, where it was confirmed that the scorecard had been received and where the indicators were explained. Applicants were not shown the scorecards as part of the intervention and were not informed about which offices were receiving the scorecards.

The scorecards were accompanied by an explanatory note stating that the office had been randomly selected to receive monthly scorecards and that the scorecards are being tested in a collaboration between a2i, the Land Reforms Board of Bangladesh, Innovations for Poverty Action (IPA), and the author. The note also explained how the performance indicators were calculated and included a phone number to call to ask questions.

C. *Randomization and Implementation Timeline*

Figure 2 depicts the randomized interventions and data collection. The scorecard intervention was carried out in two waves; each wave randomized all offices where the e-governance system had been installed at that time into either the treatment group or the control group. I conducted the first randomization in August 2018: 56 of the 112 land offices where the e-governance system had been installed were randomly selected to receive the scorecards. The first wave of scorecards started in September 2018. By April 2019, 199 additional offices had installed the e-governance system, and 99 of these offices were selected to receive the scorecards in a second randomization. The second wave of scorecards started in April 2019. The scorecards were sent out monthly until March 2020, when the COVID-19 pandemic caused the intervention to end. Supplemental Appendix B.2 provides details about the stratification of the randomization.

Additional Intervention: Peer Performance List.—To test for peer effects, a Peer Performance List was added to the scorecards for 77 randomly selected offices already receiving scorecards. The lists were added a year after the first scorecards were sent out. The list contained the percentile rankings of the two performance indicators for all 77 offices and informed them that the 76 other offices had been provided with the same information. Supplemental Appendix Figure D5 shows an example of such a list.

[11] No scorecards were sent in January 2019 due to some bureaucrats' responsibilities the previous month being shifted to the December 2018 elections instead of their regular duties.

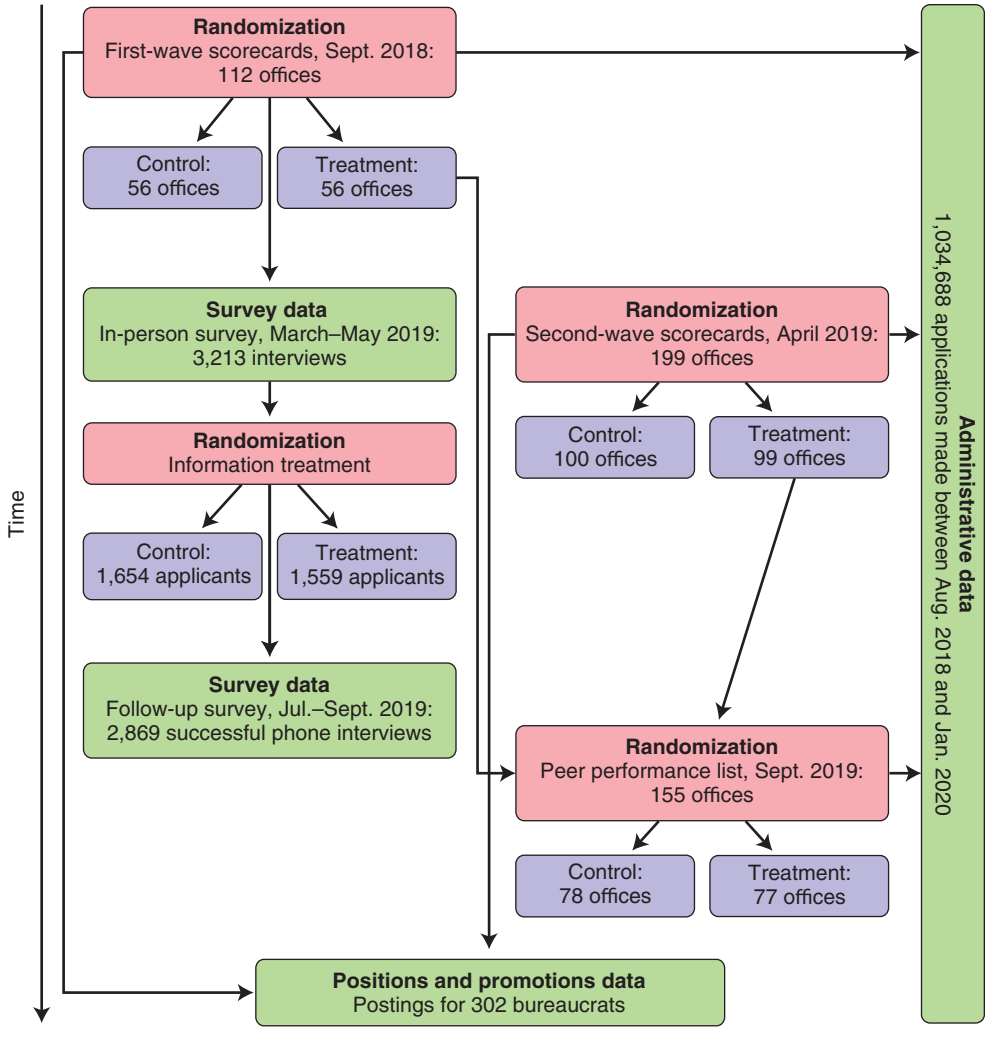

FIGURE 2. OVERVIEW OF RANDOMIZATIONS AND DATA COLLECTION

Notes: The figure displays the experiment design and data collection. The timeline is chronological from top to bottom. See discussion in Section IC.

D. *Data*[12]

I use three main data sources: administrative data from the e-governance system for all land offices in the experiment, data from a survey conducted among applicants in the offices that were part of the first randomization wave, and data from the Ministry of Public Administration on bureaucrats' designations and postings after the experiment concluded.[13] Table 1 shows summary statistics for each dataset.

[12] Mattsson (2024) provides the code used for this project, as well as the data that can be publicly shared.

[13] I also use administrative boundaries, data from the 2011 Bangladesh Census, and data from Transparency International Bangladesh as supplementary data sources (BBS 2014; Transparency International Bangladesh 2016; BBS 2023; Transparency International Bangladesh 2015).

Table 1—Summary Statistics

	Mean (1)	Median (2)	SD (3)	Observations (4)
Panel A. Application-level admin. data				
Process time $\leq$ 45 working days	0.59	1	0.49	1,034,688
Actual processing time (working days)	54	35	52	1,006,272
Processing time, incl. imputed values	61	36	69	1,034,688
Approval rate	0.67	1	0.47	1,006,265
Panel B. Monthly office-level admin. data				
Total applications	282	211	268	4,516
Applications processed	241	136	332	4,516
Apps. processed $\leq$ 45 working days	151	79	195	4,516
Apps. pending $>$ 45 working days	358	91	688	4,516
No ACL assigned	0.13	0	0.33	4,516
Female ACL	0.34	0	0.47	3,947
Panel C. Bureaucrat data				
Avg. performance percentile	54	54	16	302
Promotion by March 2020	0.68	1	0.47	302
Panel D. Applicant survey data				
Applicant age	47	47	13	2,760
Female	0.06	0	0.23	2,869
Applicant monthly income (BDT)	23,552	20,000	19,811	2,653
Applicant HH per capita expenditure (BDT)	4,396	3,400	3,579	2,869
Land value (BDT 100,000)	19	8	31	2,671
Land size (acre)	0.24	0.09	0.40	2,748
Typical payment amount (BDT)	6,718	5,000	8,416	1,802
Typical payment $>$ 0	0.75	1	0.43	1,802
Reported payment amount (BDT)	1,477	0	3,480	2,869
Reported payment $>$ 0	0.28	0	0.45	2,869

Notes: The table reports summary statistics for applications, office-months, and bureaucrats in the administrative data, as well as applicants in the survey data. Panel C only includes the bureaucrat holding the post at the start of the experiment, and the performance data are from the start of the experiment and onward. In panel D, *reported payment* amount is any payment reported by the applicant above the official fee, and *typical payment* amount is the answer to the question of how much it is "normal for a person like yourself to pay." Observations in panels A and D are inversely weighted by the number of observations in that land office. Continuous variables in the survey data are winsorized at the ninety-ninth percentile. US$/BDT $\approx$ 84.3. See discussion in Section ID.

Administrative Data from E-governance System.—The administrative data are based on 1,034,688 applications from 311 land offices (a2i and Government of Bangladesh 2020). The data contain information about the office in which the application was made, the application start date, the date it was processed, the decision to accept or reject the application, and which bureaucrat processed the application.[14] The administrative data were downloaded from the e-governance system at the beginning of each month from August 2018 until December 2020.

For the main analysis, I use administrative data for applications made from August 13, 2018, one month before the start of the intervention, until January 20, 2020. I include applications made one month before the intervention if they had not been processed by the arrival of the first scorecard, as these were partially treated. I chose

[14] Information about the applicants, such as names and phone numbers, is not available for research purposes due to privacy concerns.

to end the data 45 working days before March 26, 2020, when the COVID-19 pandemic caused a long general holiday and the end of the intervention.

The processing time is measured as the number of working days between the application start date and the date the processing was finalized. For the 2 percent of applications that had not yet been processed by December 2020, I impute the processing time by taking the mean of actual processing times that were longer than the time the application had been pending.[15] Supplemental Appendix B.5 provides more information about the administrative data.

Survey Data.—I collected two rounds of survey data from a sample of applicants who applied in the 112 offices that were part of the first randomization wave (Mattsson 2024). To create the sample of applicants, enumerators were stationed outside land offices to interview all applicants entering the office for the purpose of a land-record change, regardless of their stage in the application process. The enumerators stayed outside a specific office for at least two days and until they had completed at least 20 interviews. This first-round interview focused on the basic details of the application and applicant, the applicant's expectation for the application processing time, and the applicant's willingness to pay for faster processing.

The follow-up interviews, conducted by phone approximately three months after the initial interview, focused on the outcome of the application and bribe payments. Enumerators were not informed about which offices had received the scorecards or whether they were calling a respondent from a treatment or control office. Supplemental Appendix B.5 provides more information about the survey data.

There are two measures of bribe payments. The first is based on a question of how much the applicant thinks it is "normal for a person like yourself to pay." For the 63 percent of respondents answering this question, the amount is recorded as the variable *typical payment.* The average response is BDT 6,718 (US$80), and 73 percent of the responses were nonzero amounts.[16] The second measure is based on a series of questions about each applicant's actual payments to any government officials or agents assisting with the application. The outcome variable *reported payment* is the sum of the reported amounts. The average reported payment is BDT 1,477 (US$18), and 27 percent of respondents provided a nonzero value. Among those reporting a nonzero amount, the average amount was BDT 5,283 (US$63). Supplemental Appendix B.5 discusses the measurement of bribes in more detail.

The typical payment measure is my preferred measure of bribes, as the large number of zero responses in the reported payment measure suggests that it is an underestimate of bribes paid. However, I have no reason to believe that either of the two measures is biased differently between the treatment and control offices. Throughout the paper, I show that the main results are robust to using either of the two measures.

[15] The estimate of the scorecards' effect on the share of not-yet-processed applications is a decrease of 0.4 percentage points.

[16] To avoid extreme outliers potentially caused by enumeration errors, all continuous variables from the survey are winsorized at the ninety-ninth percentile.

Of the 3,213 applicants from the in-person interviews, 2,869 were successfully interviewed in the follow-up phone call, for a total attrition rate of 11 percent. The estimated effect of the scorecards on the attrition rate is 3 percentage points (p-value $= 0.08$). In Supplemental Appendix C.1, I discuss attrition and nonresponses in detail and use Lee bounds (Lee 2009) to show that the differential attrition is not sufficiently large to substantially affect the main findings from the survey data.

Bureaucrat Promotion Data.—I scrape data from historical versions of the Ministry of Public Administrations website for every available month between June 2014 and March 2023 using the Wayback Machine.[17] This generated a total of 384,056 bureaucrat-by-posting-by-month observations. The e-governance system's administrative data are used to determine which bureaucrat was posted to which land office at the start of the experiment. I merge the two datasets based on the names of the bureaucrats. I can then determine if the bureaucrat was promoted by the end of the experiment. Supplemental Appendix B.5 provides more information about the bureaucrats' positions data.

E. *Additional Intervention: Providing Information to Applicants*

An additional experimental intervention giving applicants information about processing times was also carried out during the first round of the survey. The motivation for this intervention was to ensure that applicants knew about the improvements in processing times. On randomly selected days, the enumerators gave applicants leaflets that described how the median processing time for all land offices had been substantially reduced over the past six months and that a new e-governance system had been implemented. The information was the same in treatment and control offices. Supplemental Appendix Figure D5 shows an English translation of the leaflet.[18]

F. *Balance of Randomization*

Panel A of Supplemental Appendix Table D1 shows balance-of-randomization tests for variables from the administrative data. To exclude all data that the scorecards could have affected, I restrict the data to applications made at least 45 working days before the start of the experiment. Applications not processed by the start of the experiment were assigned an imputed processing time based on the time they had been pending at the start of the experiment, using the imputation procedure described in Section ID. There are no statistically significant differences between scorecard and control offices before the start of the experiment. This is expected, given the random treatment assignment.

[17] https://wayback-api.archive.org

[18] Due to some noncompliance with the treatment assignment by the enumerators delivering the treatment, I use the median treatment delivered in a land office survey day as the main treatment variable. Supplemental Appendix C.8 discusses this choice and shows the robustness of the results to using alternative treatment variables.

Panel B of Supplemental Appendix Table D1 shows that the scorecards did not affect the composition of applicants or applications in the survey data. This is not a traditional balance-of-randomization test, since the treatment may have affected which applicants decided to apply and what type of applications to make. However, I find no evidence for any such changes in composition. Comparing the age and income of the applicants, the size and value of the land the applications are for, and the stages that the applications are in at the time of the first interview, there are no statistically significant differences.

II. Empirical Strategy

The experiment was preregistered, and a detailed preanalysis plan (PAP) was published on the AEA RCT Registry (Mattsson 2021).[19] The PAP describes the main analyses in the paper, but the paper also deviates from the PAP in some ways. Most importantly, the model described in the PAP is rejected by the results of the experiment. While rejecting the prespecified model is one of the results of this study, described in Section IVB and Supplemental Appendix A.2, the paper focuses on an alternative model that is consistent with the results, described in Section V. Other deviations from the PAP are reported in Supplemental Appendix B.6.

A. *Overall Effects*

To estimate the effects of the scorecards, I use the following regression specification:

$$(1) \qquad Outcome_{ait} = \alpha + \beta Treatment_i + Stratum_i + Month_t + \varepsilon_{ait},$$

where $Outcome_{ait}$ is an outcome for application a, in land office i, made in calendar month t. $Stratum_i$ are randomization stratum fixed effects. Since no randomization stratum overlaps the two randomization waves, these fixed effects also control for randomization-wave fixed effects. $Month_t$ denotes fixed effects for the month the application was made.[20] For the main results, I provide p-values testing the null hypothesis of no effect using conventional standard errors clustered at the office level, as well as p-values based on randomization inference.[21] Each observation is weighted by the inverse of the number of observations in each land office. This has three benefits. First, it makes the regression estimate the average effect of the scorecards on a land office, the unit relevant for studying changes in bureaucrat behavior.[22] Second, using these weights, the analyses of the administrative data and the

[19] www.socialscienceregistry.org/trials/3232

[20] In the survey data, consistent with the cleaning of other continuous variables, the application month variable is winsorized at November 2018 so that all application dates in or before November 2018 take the same value. A separate indicator variable controls for missing start-date values.

[21] The randomization inference is implemented using the Stata command *randcmd*, and the reported p-value is from the randomization t-test calculated using 9,999 iterations (Young 2019).

[22] Supplemental Appendix Table D2 confirms that the results are similar when collapsing the data to the office level and measuring the effects on the offices' mean of each outcome.

survey data estimate the same effect.[23] Third, the weighting improves the estimates' precision by weighting each cluster equally in the analysis.[24]

The two additional randomized interventions—the peer performance lists and the information intervention to applicants—are excluded from the main specification. Section IVB and Supplemental Appendix C.6 discuss the effects of these interventions on bribe payments and processing times. These sections also show that neither of the two additional interventions has substantial interaction effects with the scorecard treatment, validating my approach of analyzing the scorecard treatment separately.

B. *Heterogeneous Effects by Baseline Office Performance*

A reoccurring result in the literature on performance feedback and monitoring interventions is that positive effects are driven by low performers and that high performers display smaller and sometimes even negative effects. This is shown by Allcott (2011); Ayres, Raseman, and Shih (2013); and Byrne, Nauze, and Martin (2018) for electricity consumption in the United States; Barrera-Osorio et al. (2020) for students in Colombia; and Dodge et al. (2021) for Indian bureaucrats.

Following this literature, I separate offices by their baseline performance and estimate the effect of the scorecards separately for offices performing above and below the median at baseline. As per my PAP, I define baseline performance as the average of the two performance indicators' percentile rankings in the first month of treatment and classify all offices into *overperformers* (above the median baseline performance) and *underperformers* (below the median baseline performance).[25] The above- and below-median baseline performance classification was the only heterogeneity test based on office characteristics specified in the PAP.[26] Since the classification of offices uses data only from before the first scorecard was delivered, it is not affected by the treatment. Supplemental Appendix Table D3 shows that the results are robust to alternative measures of baseline performance.

I use the following regression specification to estimate the effect of the scorecards on the two types of offices separately:

$$(2)\quad y_{ait} = \alpha + \beta_1 Treatment_i \times Overperform_i + \beta_2 Treatment_i \times Underperform_i + \gamma Overperform_i + Stratum_i + Month_t + \varepsilon_{ait},$$

[23] With uniform weights, the administrative data analysis estimates the average effect on applications in the e-governance system. However, the survey data estimate the average effect among surveyed applicants, which is roughly equal to the average effect on a land office, as most land offices have a similar number of surveyed applicants.

[24] As illustrated in Supplemental Appendix Figure D4, panel A, some land offices receive more applications than others. This causes 57 percent of the administrative data sample to come from the 25 percent largest offices, while only 6 percent come from the 25 percent smallest offices. For a discussion of this weighting, see https://blogs.worldbank.org/impactevaluations/different-sized-baskets-fruit-how-unequally-sized-clusters-can-lead-your-power.

[25] I classify 112 offices in the first randomization wave into over- and underperformers by comparing them to the median performance among these 112 offices. For the offices in the second randomization wave, I compare them to the median performance of all 311 offices in the experiment at the time of their first scorecard. This makes the *overperformer* and *underperformer* classifications correspond to the relative performance presented in the first scorecards. However, it also means that slightly more than half (53 percent) of all offices are classified as overperformers.

[26] The two other prespecified heterogeneity tests were based on the date of application and the application processing time and appear in Supplemental Appendix Figure D6 and Supplemental Appendix Table D5, respectively.

where β_1 is the estimated effect of the scorecards for offices overperforming at baseline, β_2 is the effect for offices underperforming at baseline, and γ is the difference between overperforming and underperforming offices in the control group.

To test the hypothesis that the treatment had the same effect on offices overperforming and underperforming at baseline, I use a regression almost identical to the regression described in equation (2), but the first treatment variable is not interacted with $Overperform_i$. I then test the hypothesis that the coefficient on $Treatment_i \times Underperform_i$ is zero. This test's p-value is reported as "p-value: subgroup diff." in the regression tables.

Supplemental Appendix Table D4 compares the over- and underperforming offices and their subdistricts. There is a clear difference in performance, even after the baseline period. Overperforming offices processed 21 percentage point (43 percent) more of the applications on time. Overperforming offices collected fewer bribes than underperforming offices, despite the intervention increasing bribe payments among overperforming offices. Overperforming offices were also more likely to be female and approve a higher share of applications. Apart from these differences, the offices were similar in many observable characteristics such as applications received per month, lacking a bureaucrat assigned to the office, subdistrict population, area, and share of the labor force in agriculture.

III. Results: Effects of the Scorecards

A. *Effect on Processing Times*

Panel A of Table 2 shows that the scorecards increased the number of applications processed within the government time limit and improved processing times overall using the empirical specification in equation (1). Column 1 shows the estimated effect of the scorecards on a binary variable indicating whether the application was processed within the 45-working-day time limit. The scorecards increased applications processed within the limit by 6 percentage points (11 percent). Column 2 shows that the scorecards led to a 13 percent reduction in overall processing times by estimating the effect on the natural logarithm of the processing time.[27]

For column 3, I create a *time index* of the two outcomes used in columns 1 and 2.[28] The estimated effect of the scorecards on the time index is 0.13 standard deviations (conventional p-value $= 0.028$; randomization inference p-value $= 0.037$).[29]

Heterogeneity of Effect on Processing Times by Baseline Office Performance.— Panel B of Table 2 uses the empirical strategy from Section IIB to show that the scorecards' effect on processing times is driven by offices that were underperforming

[27] The exact effect is -12.6 log points, which is equivalent to an 11.9 percent decrease. For simplicity, I will describe log point changes as percentage changes throughout the paper.

[28] I created the index by first taking the negative of the log processing time so that a higher value indicates better performance, then recasting the two outcome variables as standard deviations away from the control group mean, and finally, taking the sum of the two standard deviations and rescaling them so that the index has a standard deviation of 1 in the control group.

[29] Supplemental Appendix C.2 shows similar results using the survey data.

Table 2—Scorecards' Effect on Application Processing Times

	≤ 45 working days (1)	ln(working days) (2)	Time index (3)
Panel A. Overall effect			
Scorecard	0.060	−0.126	0.131
	(0.027)	(0.059)	(0.059)
Panel B. Heterogeneous effects			
Scorecard × overperform	0.011	−0.043	0.034
	(0.037)	(0.080)	(0.080)
Scorecard × underperform	0.115	−0.218	0.238
	(0.040)	(0.088)	(0.088)
Overperform baseline	0.194	−0.313	0.374
	(0.050)	(0.107)	(0.108)
p-value: subgroup diff.	0.064	0.149	0.092
Start-month and stratum FEs	Yes	Yes	Yes
Observations	1,034,688	1,034,688	1,034,688
Clusters	311	311	311
Overperformers: control mean	0.69	49.57	0.27
Underperformers: control mean	0.42	80.13	−0.30

Notes: The table reports the effect of the scorecards on the speed of application processing. Column 1 shows the effect on applications processed within the time limit. Column 2 shows the effect on the log of processing time. Column 3 shows the effect on an index combining the two outcome variables. The data contain all applications made between one month before the start of the experiment and 45 working days before the experiment ended. Panel B reports the effect of the scorecards separately for offices with above- and below-median baseline performance. Observations are inversely weighted by the number of observations in that land office. Standard errors are clustered at the office level. See discussion in Section IIIC.

at baseline. Column 3 shows that, for offices that were underperforming at baseline, the estimated effect on the time index is an increase of 0.24 standard deviations (conventional p-value $= 0.007$; randomization inference p-value $= 0.012$).[30] For offices overperforming at baseline, the effect is just 0.03 standard deviations (p-value for the difference in effects is 0.092).

Supplemental Appendix Table D6 shows heterogeneous effects based on what thumbs-up and thumbs-down symbols were shown on the first scorecard. The table shows that the scorecards with thumbs-down symbols were particularly effective in improving processing times and that this differential effect remains even after including linear controls for the two rankings displayed on the scorecards and interacting these rankings with the scorecard treatment. This is suggestive evidence that the symbols made the scorecards clearer and more salient. Supplemental Appendix C.5 discusses the effects of the symbols in more detail.

Effect on the Distribution of Processing Times.—Figure 3 shows the effect using overlaid histograms and cumulative distribution functions (CDFs) of minimally processed data on processing times. Figure 3, panel A and panel B show that in the treatment offices, more applications were processed within the 45-working-day

[30] The randomization p-value is 0.023 when using the Westfall-Young multiple hypothesis testing method to adjust for testing two hypotheses, one for overperformers and one for underperformers.

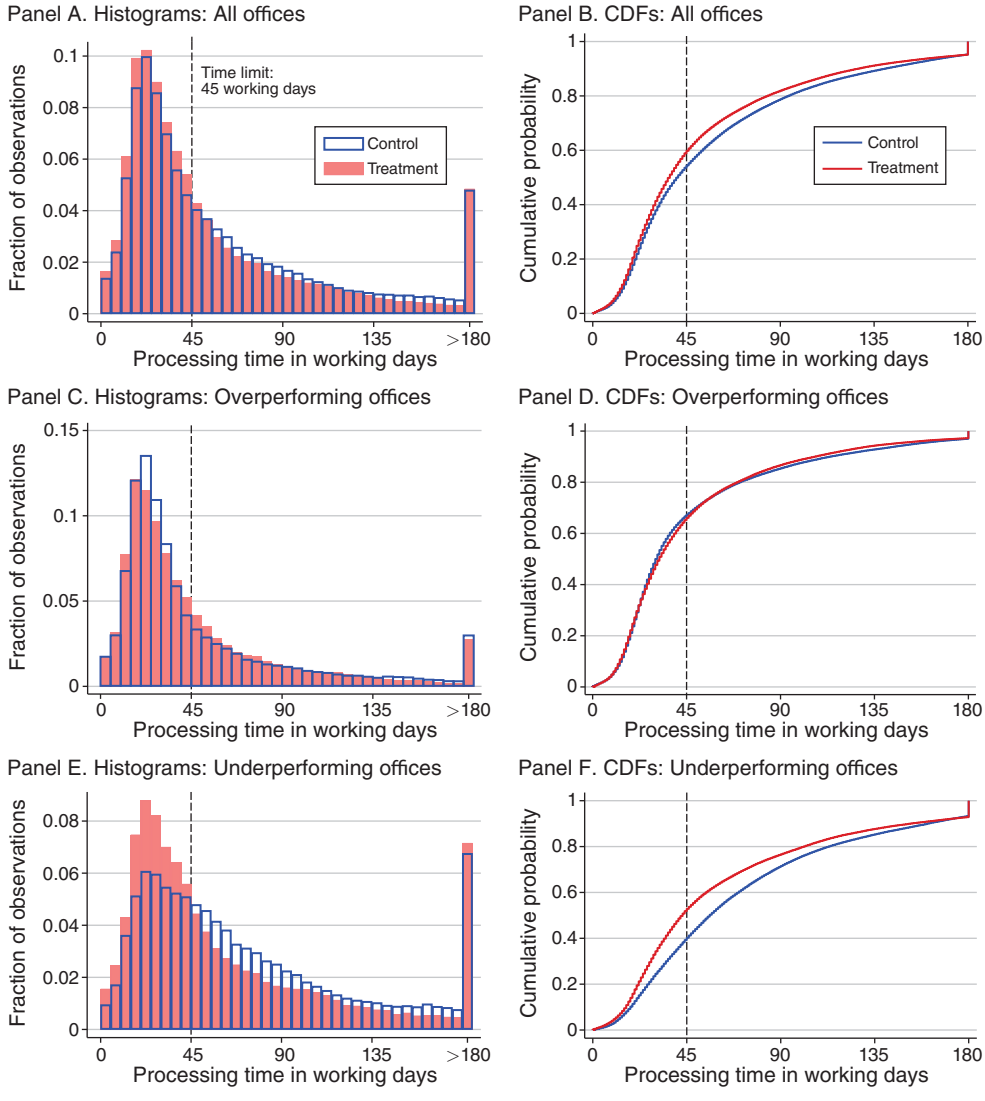

FIGURE 3. PROCESSING TIME DISTRIBUTIONS BY TREATMENT STATUS

Notes: The figure displays histograms and CDFs of processing times for the treatment and control groups separately. Processing times are top coded at 180 working days. Panels A and B use data from all offices. Panels C and D use data from offices overperforming at baseline. Panels E and F use data from offices underperforming at baseline. The vertical dashed lines represent the 45-working-day time limit. See discussion in Section IIIA.

time limit. The effect is relatively evenly spread over the whole distribution, with no substantial bunching just before the 45-day limit.[31] Processing times that are reduced in frequency by the scorecards are in the whole span from 55 working days and up. This is reasonable, given that the scorecards emphasize both processing

[31] This is to be expected given that the process for approving an application is relatively long and depends on several individuals, as described in Section IA. Thus, even if the bureaucrat cares only about maximizing the share of applications processed within 45 working days, the processing time target has to be lower than 45 working days.

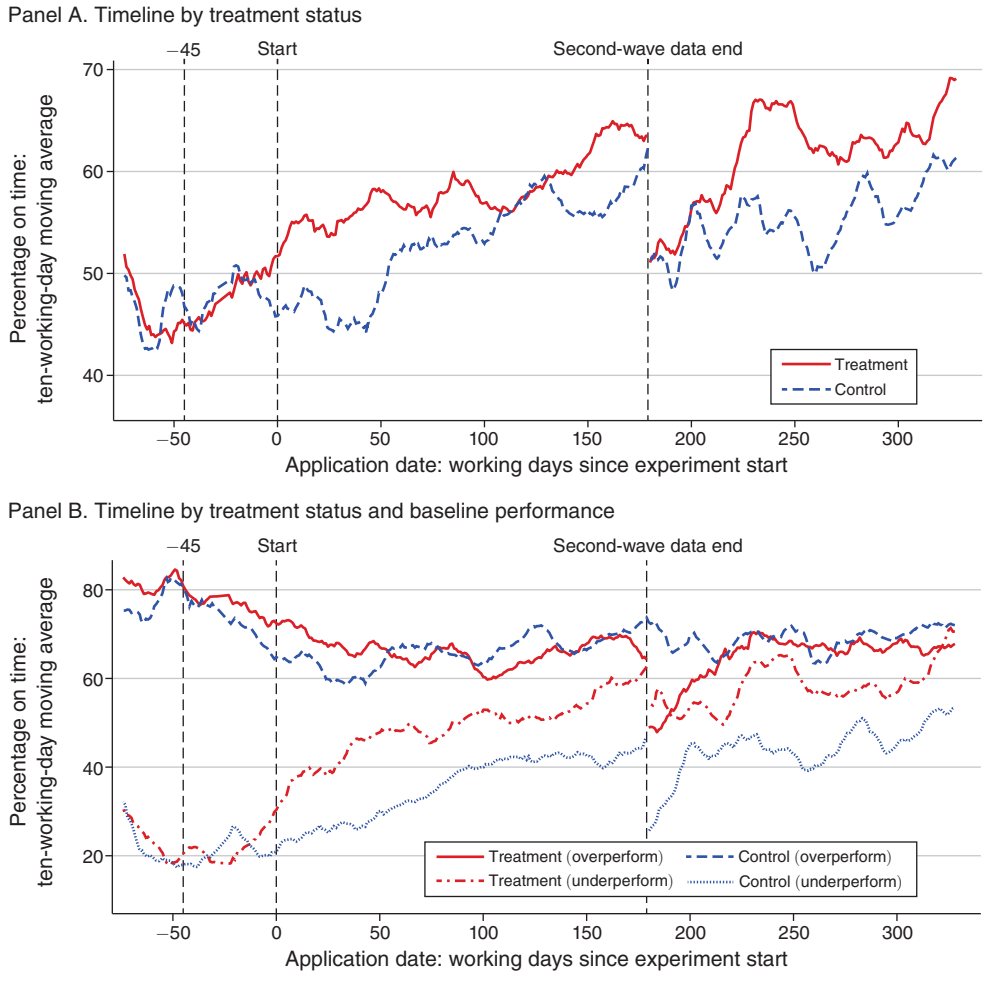

FIGURE 4. APPLICATIONS PROCESSED WITHIN TIME LIMIT OVER TIME

Notes: The figure displays ten-working-day moving averages of the share of applications processed within the time limit. The data are arranged by application start date relative to the start of the experiment. The first and second vertical lines represent the date 45 working days before the first scorecard and the date of the first scorecard, respectively. Applications made between these dates are partially treated. The third vertical line represents the end of the data from the second randomization wave. See discussion in Section IIIA.

applications within the 45-working-day limit and reducing the number of applications pending beyond the limit. Figure 3, panels C to F show the changes in the distributions for over- and underperforming offices separately. Supplemental Appendix Figure D7 shows the effect on the distribution of offices' share of applications processed on time.

Effect over Time.—It is possible that the initial effect of the scorecards is different from the long-run effect. If there was a substantial novelty effect, we would expect to see the difference between the treatment and control groups decline over time. Figure 4 and Supplemental Appendix Figure D6 show no pattern of a declining effect over the 16 months of the intervention, though the size of the effect varies between different periods.

Figure 4 shows the ten-working-day moving average of the share of applications processed within the time limit by application date relative to the start of the experiment. The first and second vertical lines indicate the date 45 working days before the first scorecards were sent and the day of sending them, respectively. Applications made between these dates were partially affected by the scorecards, as the bureaucrats received the scorecards while processing these applications. The data from the offices in the first randomization wave cover a longer time period relative to the start of the experiment in these offices than the data from the offices in the second randomization wave. The third vertical line marks where the data from the second randomization wave end. Starting with applications made just before the first scorecards were sent, we see that the treatment group processes more applications on time relative to the control group. With a few short exceptions, the treatment offices continue to process a higher share of applications within the time limit until the end of the experiment.[32] Figure 4, panel B shows that the heterogeneity in the effect between offices over- and underperforming offices at baseline is persistent over time. It also shows that the treated offices underperforming at baseline catch up with offices overperforming at baseline, but that they do not overtake these offices.

B. *Effect on Bribe Payments*

Panel A of Table 3 shows that the scorecards did not lead to a decrease in bribe payments. Instead, the estimated effect on bribes is positive, though this increase is not statistically significant. As described in Section ID, data on bribe payments were collected using two separate survey questions. The first question asked about the typical bribe payment "for a person like yourself." When this measure is used, the column is marked as *typical*. The second question asked about each payment made by the applicant. When this measure is used, the column is marked *reported*.

Columns 1 and 2 of Table 3 show the effect on the amount of bribes paid. Column 1 shows that the effect on the perceived typical payment was BDT 940 (US$11), a 15 percent increase (conventional p-value $= 0.130$; randomization inference p-value $= 0.159$). The lower bound of the confidence interval is a decrease of 4 percent, ruling out a meaningful decrease. Column 2 estimates that the scorecards increased reported bribe payments by BDT 297, a 23 percent increase.

Columns 3–6 of Table 3 disaggregate the effect into extensive and intensive margin effects. Columns 3 and 4 show that there is no effect on the propensity to report a nonzero bribe. This can be interpreted as the scorecards having no effect on the extensive margin of bribe payments. Another interpretation is that the intervention did not affect applicants' willingness to talk about bribe payments in the survey. To measure the effect on the intensive margin, that is, the amount paid among those paying some bribe, columns 5 and 6 restrict the sample to those who gave a strictly positive (i.e., nonzero) response to the bribe payment questions. On the intensive

[32] Estimating the effect on the time index for applications made between March 26, 2020 (the end of the intervention), and September 28, 2020 (45 working days before the end of my data), yields an estimate of 0.07 standard deviations. This suggests that there is only a small amount of persistence in the effect of the scorecards when they are no longer sent. However, this estimate should be interpreted carefully as the COVID-19 pandemic drastically changed the operating conditions for the bureaucrats during this period.

Table 3—Scorecards' Effect on Bribe Payments

	Amount		Any bribe		Amount if > 0	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Overall effect						
Scorecard	939	297	−0.02	−0.00	1,491	1,169
	(616)	(182)	(0.02)	(0.02)	(762)	(452)
Panel B. Heterogeneous effects						
Scorecard × Overperform	2,069	629	0.03	0.04	2,467	1,743
	(765)	(233)	(0.03)	(0.03)	(951)	(622)
Scorecard × Underperform	−99	42	−0.06	−0.04	622	797
	(957)	(259)	(0.03)	(0.03)	(1,196)	(646)
Overperform baseline	−1,833	−814	−0.09	−0.09	−1,416	−1,307
	(977)	(294)	(0.04)	(0.03)	(1,173)	(729)
p-value: subgroup diff.	0.09	0.10	0.08	0.07	0.24	0.30
Start-month and stratum FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,802	2,869	1,802	2,869	1,324	779
Clusters	112	112	112	112	112	111
Overperf. control mean	5,455	944	0.717	0.229	7,606	4,131
Underperf. control mean	6,726	1,578	0.788	0.316	8,535	4,992
Bribe measure	Typical	Reported	Typical	Reported	Typical	Reported

Notes: The table reports the effect of the scorecards on bribe payments. Columns 1, 3, and 5 show the effect of the scorecards on the response to the question about the value of a typical payment by "a person like yourself." Columns 2, 4, and 6 show the effect on the response to the question about actual payments to government officials or agents. Columns 3 and 4 show the effect on the percentage of nonzero responses to the questions (extensive margin). Columns 5 and 6 show the effect among applicants who reported a nonzero bribe (intensive margin). Panel B reports the effect of the scorecards separately for offices with above- and below-median baseline performance. All monetary amounts are in BDT. US$/BDT $\approx$ 84.3. All continuous variables are winsorized at the ninety-ninth percentile. Standard errors are clustered at the office level. Observations are inversely weighted by the number of observations in that land office. See discussion in Section IIIB.

margin, bribe payments increased by 18 percent for typical payments and 25 percent for reported payments. Supplemental Appendix Figure D8 shows the differences in the distributions of typical payments between treatment and control groups.

Heterogeneity of Effect on Bribe Payments by Baseline Office Performance.— Panel B of Table 3 uses the empirical strategy from Section IIB to show that the positive effect on bribe payments is entirely driven by the offices that were overperforming at the start of the experiment. Column 1 of panel B shows that the effect of the scorecards on estimated typical bribe payments among offices overperforming at baseline is an increase of BDT 2,069, equivalent to 38 percent (conventional p-value $= 0.008$; randomization inference p-value $= 0.016$).[33] Supplemental Appendix Table D7 shows that the positive effect on bribes for overperforming offices is spread out across different stated reasons for paying the bribe, without any one reason driving the effect.

The effect on offices underperforming at baseline is close to zero (p-values for the differences in effects between over- and underperformers are 0.086 and 0.098).

[33] The randomization inference p-value is 0.032 when using the Westfall-Young multiple hypothesis testing method to adjust for testing two hypotheses.

However, due to the relatively imprecise measurements, I cannot rule out meaningful negative or positive effects for underperformers; the upper bound of the 95 percent confidence interval for typical payments is a 26 percent increase, while the lower bound is a 29 percent decrease.

Supplemental Appendix Tables D8 and D9 show that the heterogeneous effects on processing times and bribes by baseline performance remain similar after controlling for other office baseline characteristics interacted with the treatment. Furthermore, Supplemental Appendix Table D4 shows that over- and underperforming offices are similar along most several characteristics except speed of service delivery and bribe taking. However, the association between the effect size and baseline performance should still not be interpreted as a causal relationship. Baseline performance is not randomly assigned, and it is plausible that unobserved office characteristics are associated with both performance and the treatment effect size. Instead, I interpret the results as describing which type of offices reacted to the scorecards.

Even without a causal interpretation, the heterogeneity in the results is surprising: Overperforming offices did not change their behavior in terms of processing times, but among these offices, bribe payments increased. Sections IV and V investigate this further.

C. *Other Unintended Consequences*

A common problem with quantitative performance measures is that they often lead to gaming of the measures or other unintended consequences (Finan, Olken, and Pande 2017). In Supplemental Appendix C.4, I test for four such potential unintended consequences. First, if bureaucrats allow fewer applicants to start applications, then their scorecards may improve, provided that the lower number of applications helps them process a larger share of the applications within the time limit. Second, if bureaucrats allowed applications selectively, such that the average application was easier to process within the time limit, then their scorecards may improve. Third, the scorecards may lead bureaucrats to make worse decisions regarding accepting or rejecting applications. Fourth, bureaucrats may divert attention from applications not made in the e-governance system because those applications do not count toward the scorecards. I find no evidence for large unintended consequences except for suggestive evidence of a higher incorrect rejection rate in offices overperforming at baseline, potentially as a response to applicants not being willing to pay the new higher bribes.

D. *Robustness Tests*

The main results for the effects of the scorecards on processing times and bribes, as well as the heterogeneity in these effects, are robust to a range of alternative specifications. Supplemental Appendix Tables D10 and D11 show the results using various combinations of controls, weights, and winsorizations of the bribe amounts, as well as including only bribes given directly to government officials while ignoring fees paid to agents. All alternatives to the main estimates are of the same sign and of similar magnitude, but some of them are not statistically significant. Supplemental

Appendix Tables D12 and D13 show the results when controlling for the number of applications received by the office. Again, the main results remain similar. Supplemental Appendix Table D14 shows the effects, estimated at the office-month level, on the number of applications processed within 45 working days and the number of applications pending beyond 45 working days, as well as those figures' corresponding percentile rankings. The point estimates suggest that the scorecards improved all four of these outcome variables, driven by improvements in offices underperforming at baseline, but only the results for underperforming offices are statistically significant. Supplemental Appendix Table D2 shows that the results are similar when measuring the effects on the offices' mean of each outcome. Supplemental Appendix Table D15 shows that the estimated effect on processing times is robust to different functional form assumptions and different imputation techniques for applications that were not yet processed.

Supplemental Appendix Table D3 shows that the heterogeneity found in the effects is robust to three alternative measures to the prespecified measure of baseline performance, including using quartiles and a continuous measure of baseline performance. Supplemental Appendix Table D16 shows the results are robust to restricting the sample to applications made before the survey took place and applications made in offices where there was no survey, showing that the survey did not substantially alter the treatment effect.

IV. Mechanisms and Effect on Promotions

A. *Mechanisms for the Effect on Processing Times*

The scorecards could improve the performance of bureaucrats by providing information to their supervisors. This could improve the supervisors' ability to incentivize the bureaucrats by, for example, facilitating faster promotions for those bureaucrats with good scorecards while delaying promotions for poorly performing bureaucrats.

Table 4 tests this hypothesis and shows suggestive evidence that the scorecards delayed promotions for bureaucrats underperforming at baseline while not affecting the promotion for overperforming bureaucrats. The data are cross-sectional, and each observation is a bureaucrat assigned to a land office at the start of the experiment.[34] Column 1 uses the bureaucrat-level data to confirm the results from Section IIIA: that the scorecards improved the performance of the bureaucrats and that this effect is driven by underperforming bureaucrats. The outcome variable is the bureaucrats' average performance percentile across the two performance indicators after the start of the experiment.

Column 2 of Table 4 shows a negative point estimate of 5 percentage points (-7 percent) for the effect of the scorecards on the share of bureaucrats promoted, but the estimate is not statistically significant.[35] The negative effect is driven by bureaucrats underperforming at baseline for whom the effect is a 17 percentage

[34] If no bureaucrat was assigned to the office at the start of the experiment, I use the first bureaucrat assigned to the office within the first year after the start of the experiment.

[35] Promotion is defined as holding any position with a higher rank than assistant commissioner.

TABLE 4—SCORECARDS' EFFECTS ON BUREAUCRATS' PERFORMANCE AND PROMOTION

	Avg. performance rank (1)	Promoted (2)
Panel A. Overall effect		
Scorecard	3.413	−0.049
	(1.421)	(0.054)
Panel B. Heterogeneous effects		
Scorecard × Overperform	0.723	0.056
	(1.935)	(0.074)
Scorecard × Underperform	6.439	−0.171
	(1.847)	(0.077)
Overperform baseline	15.709	−0.179
	(2.432)	(0.093)
p-value: subgroup diff.	0.008	0.036
Stratum FE	Yes	Yes
Observations	302	302
Control mean	53.03	0.70

Notes: The table reports the effect of the scorecards on bureaucrats' performance and promotions. The data are cross-sectional, and each observation is a bureaucrat assigned to a land office when the experiment started. Column 1 shows the effect on bureaucrats' average monthly performance percentile in terms of the two performance indicators included in the scorecards. Column 2 shows the effect of the scorecards on being promoted by the end of the experiment (March 2020). Heteroskedasticity robust standard errors. See discussion in Section IVA.

point decrease (−25 percent, conventional p-value 0.027; randomization inference p-value $= 0.029$).[36] Although the estimates are imprecise, they suggest that receiving a negative performance scorecard is negative for a bureaucrat's promotion prospects, while receiving a positive scorecard has only a small positive or zero effect on promotion prospects.

An additional potential mechanism for the effects of the scorecards is that bureaucrats may change their behavior due to receiving the scorecards themselves. For bureaucrats, receiving information about their processing times each month may increase this information's salience, causing it to be more important for their personal sense of pride in their work. Since the scorecards were sent to both bureaucrats and their supervisors, I cannot separately estimate the importance of these two mechanisms, and I refer to them collectively as *reputational concerns*.

Information flows about performance between bureaucrats at the same level in the organizational hierarchy may create an additional incentive for improved performance (Mas and Moretti 2009; Blader, Gartenberg, and Prat 2020). However, Supplemental Appendix Table D17 shows that there is no substantial effect of the peer performance list (described in Section IC) on processing times. This suggests that in this setting, information flows between bureaucrats at the same level do not

[36] The randomization p-value is 0.057 when using the Westfall-Young multiple-hypothesis testing method to adjust for testing two hypotheses.

have an effect beyond the effect of sending the scorecards to the bureaucrats and their supervisors. Supplemental Appendix C.6 provides more details about the results.

B. *Does Speeding Up Service Delivery Reduce Corruption?*

The experiment is testing a typical policy prescription stemming from the view that faster service delivery reduces corruption, namely improved performance monitoring of service delivery speed.[37] However, the results in Section III show that the scorecards did not reduce bribes, despite improving processing times. This is true even for the offices that were underperforming at baseline and improved their processing time the most. Furthermore, Supplemental Appendix Table D7 also shows that the scorecard did not decrease the bribes that applicants stated were for the purpose of increasing the speed of processing.

One potential reason for the lack of a negative effect on bribe payments could be that the information about the improvement in processing times had not yet been disseminated among applicants. The information treatment was designed to test this hypothesis by informing applicants about improvements in processing times, as described in Section IE.[38] However, Supplemental Appendix Table D18 shows that the information treatment did not affect bribes, neither by itself nor in combination with the scorecards, despite reducing the applicants' expected processing times.

V. Model of Bureaucrat Behavior

In this section, I outline my model explaining the results of the experiment. Supplemental Appendix A.1 presents the formal model.

A. *Model Setup*

In the model, bureaucrats get utility from reputation and bribe money, while effort causes disutility. A reputation term represents reasons why the bureaucrats care about what their supervisors think of them, as well as psychological reasons that are internal to the bureaucrats (i.e., both mechanisms described in Section IVA). The reputation term is a function of bribe money taken and visible job performance—in this case, processing times. Effort is needed to improve performance. Bribe money has decreasing marginal utility, while effort has increasing marginal disutility. There is decreasing marginal utility from reputation in performance and increasing marginal disutility through the reputation mechanism in bribe money taken. The scorecards are modeled as increasing the importance of processing times for the bureaucrats' reputations by increasing the visibility of performance.

[37] See Supplemental Appendix A.2 for a discussion of the literature and a set of empirical tests rejecting one particular model of how the processing times and corruption are related.

[38] Supplemental Appendix C.8 discusses the noncompliance with the treatment assignment in the delivery of this intervention and shows the robustness of the results to alternative definitions of the treatment variable.

I assume that bribe money and visible performance are complements in generating reputation, or equivalently, that honesty (the absence of bribe taking) and performance are substitutes. In terms of bureaucrats' career prospects, this assumption is based on the idea that some corruption is acceptable as long as bureaucrats are performing their duties well and that poorly performing bureaucrats still have a good career in the bureaucracy, as long as they follow the official rules. If bureaucrats are both corrupt and poorly performing, this could endanger their careers, while honest, high-performing bureaucrats still can't be promoted much faster than their colleagues, because most promotions are based on seniority. This is consistent with the empirical patterns discussed in Supplemental Appendix B.1 and the results on bureaucrats' promotions in Table 4, where negative performance scorecards reduced promotions, while positive scorecards did not change promotion probabilities.[39]

I assume that bureaucrats differ only in how much they value their reputation. This could be because of differences in the valuation of future career prospects or differences in intrinsic motivation to be honest and perform well.[40] These differences are what generates over- and underperforming bureaucrats.

Finally, I assume that the applicants simply pay the bribe amount that the bureaucrats are demanding. In other words, applicants' willingness to pay for the service is not a binding constraint. Instead, what determines the amount of bribes in the model is the bureaucrat's trade-off between bribe money and reputational concerns.

B. *Model Predictions*

The theoretical model has two main testable predictions. Figure 5 depicts the model's predictions compared to the empirical results. Supplemental Appendix A.1 provides the formal derivations and statements of the predictions.

Effects of the Scorecards on Processing Times.—The first set of predictions is for the effect of the scorecards on processing times. The scorecards have two effects on processing times, an *incentive effect* and a *reputation-level effect*. These effects are akin to the substitution and income effects from a wage increase in a labor supply model. The incentive effect leads to increased effort and reduced processing times for all bureaucrats. This is because the scorecards increase the visibility of the bureaucrats' performance and, therefore, the marginal effect it has on utility.

The direction of the reputation-level effect depends on if the scorecard increases or decreases the reputation of the bureaucrat. For overperforming bureaucrats, the reputation-level effect is negative because the scorecards increase their reputation by making their good performance more visible. Since reputation has decreasing marginal utility, this decreases the marginal utility of reputation and reduces the optimal

[39] An alternative rationale for this assumption is that bureaucrats feel a strong sense of shame if they are both corrupt and low performing. However, if their visual performance improves, they feel "licensed" to take more bribes without shame and vice versa. This would be consistent with the literature on moral licensing (Sachdeva, Iliev, and Medin 2009; Clot, Grolleau, and Ibanez 2018).

[40] Prendergast (2007) and Hanna and Wang (2017) provide evidence that differences in the intrinsic motivations of government officials can be important for public service delivery and corruption. Bertrand et al. (2020) provide evidence for the importance of differences in future career prospects.

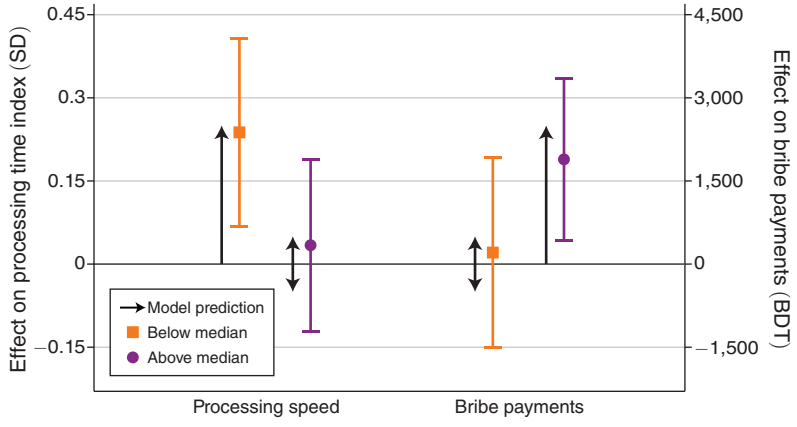

FIGURE 5. MODEL PREDICTIONS AND EMPIRICAL RESULTS

Notes: The figure presents the heterogeneous effects of the scorecards from Tables 2 and 3 compared to the model's predictions from Section VB. The arrows represent the direction of the effect predicted by the model. The empirical estimates show the estimated effects for offices performing below and above the median at baseline. The effects on processing speed are measured on the left *y*-axis in standard deviations of the time index constructed using the variables for whether the application was processed on time and the log of the overall processing time. The effects on bribe payments are measured on the right *y*-axis in BDT. US$/BDT $\approx$ 84.3. The bribe data come from the responses to the question about how much it is "normal for a person like yourself to pay." The 95 percent confidence intervals are constructed using standard errors clustered at the office level. See discussion in Section VB.

amount of effort overperforming bureaucrats exert to process applications on time. For underperforming bureaucrats, the reputation-level effect is positive because the scorecards highlight their poor performance and lower their reputation level.

The predicted effect of the scorecards for the over- and underperforming bureaucrats can therefore be asymmetrical. For overperforming bureaucrats, the incentive effect and reputation-level effect have opposite directions, and the overall effect on processing times depends on which of the two effects is stronger. For underperforming bureaucrats, the incentive effect and reputation-level effect move in the same direction, and the model predicts that the scorecards will improve processing times.

PREDICTION 1: *Scorecards improve processing times for underperforming bureaucrats.*

Figure 5 shows how the model's prediction is consistent with the scorecards improving processing times for offices underperforming at baseline, while the effect is close to zero for offices overperforming at baseline.

Effects of the Scorecards on Bribes.—The second prediction relates to the effect of the scorecards on bribes taken from applicants. In the model, the bureaucrats can increase bribes by simply asking applicants for more money. What constrains bureaucrats from extracting more bribes is the negative marginal effect it has on their reputation. When the scorecards improve overperforming bureaucrats' reputations, the marginal effect bribes have on utility through the reputation channel

becomes less negative. This leads to an increase in bribes taken by overperforming bureaucrats when they receive the scorecards.

Underperforming bureaucrats suffer an initial decrease in reputation from the information provided, but since effort increases in response to the scorecards, the overall effect on reputation could be positive or negative. Therefore, the effect of the scorecards on bribes is ambiguous for underperforming bureaucrats. This explains the asymmetry in the model's predictions for over- and underperforming bureaucrats.

PREDICTION 2: *Scorecards increase bribes for overperforming bureaucrats.*

Figure 5 shows how the model's prediction is consistent with the scorecards increasing bribes paid in offices overperforming at baseline, while the effect is close to zero for offices underperforming at baseline.

C. *Alternative Explanations*

Supplemental Appendix C.7 discusses five potential alternative explanations for the experimental results. First, the scorecards may have increased the opportunity cost of bureaucrats' time or changed bureaucrats' perceptions of what is important to their superiors toward a view where the speed of public service delivery matters more, perhaps at the expense of the importance placed on limiting corruption. These explanations cannot explain the heterogeneity of the results, because it is not the offices improving their processing times (those underperforming at baseline) that increases the amount of bribes paid (those overperforming at baseline). However, the heterogeneity in the results between over- and underperforming offices is only statistically significant at the 10 percent level. Thus, it is possible that these mechanisms were involved in generating the results.

Second, the scorecards may have increased transfers of overperforming bureaucrats away from the ACL position, and this may have caused the increase in bribe payments among offices overperforming at baseline. This is inconsistent with the estimated effects on bureaucrat transfers in Supplemental Appendix Table D19 being close to zero. Third, the supervisors may have shifted monitoring attention away from overperforming offices, leading to increased bribes. This explanation is dependent on an asymmetry where an increase in the monitoring attention on underperforming offices either did not occur or did not decrease bribes there. Fourth, the supervisors' reputation level may have increased, causing them to demand higher bribes. This is unlikely because land-record changes constitute a much smaller share of supervisors' responsibility than they do for the bureaucrats. Fifth, the bureaucrats may have used the scorecards in negotiations with applicants as "proof" that they can process the application quickly. This is not consistent with applicants in overperforming offices' expectations of processing times not improving, shown in Supplemental Appendix Table D20. It is also inconsistent with the lack of an effect from the information treatment on bribe payments, shown in Supplemental Appendix Table D18.

Above, I have shown that my model is consistent with the empirical results from the experiment and that for each alternative explanation, there is at least some

empirical pattern that cannot be explained by that mechanism alone. Therefore, I present the mechanism in my model as the most plausible explanation for the observed overall results. However, it is possible that some of these alternative explanations also affected the results.

VI. Conclusion

I have shown that an information system—providing information about delays to the responsible bureaucrats and their supervisors—can improve the delivery speed of an important public service. This effect is present despite the absence of explicit performance incentives, and it is persistent, at least over 16 months. The system, made possible by an underlying e-governance system, is an example of how new technologies are creating opportunities for improved management in the public sector.

To create a rough estimate for the value of the improved processing times, I multiply the applicants' average stated valuation of having their application processed one day faster by the reduction in the total number of processing days due to the scorecards. For the 155 offices receiving the treatment, the approximate value is US$9.7 million per year. See Supplemental Appendix C.9 for details.

This value should be interpreted carefully—as it relies heavily on the value stated by the applicants. However, the number is more than two orders of magnitude larger than the cost to implement the scorecards, which was approximately US$40,000 per year. The overall welfare effect of the intervention becomes less clear when taking into account the effect on bribe payments. Multiplying the effect of the scorecards on typical payments with the number of applications in the treatment area results in an estimate of the effect on total bribes paid of US$6.6 million per year.[41] Furthermore, the scorecards have a negative but not statistically significant effect on stated satisfaction. Overall, there is no strong evidence that the scorecards had either a positive or a negative effect on average applicant welfare.

More than half of Bangladesh's land offices took part in the experiment, making it plausible that the results are externally valid within Bangladesh (Muralidharan and Niehaus 2017). However, while I designed the scorecards in collaboration with the government, they were produced and distributed by a nonprofit research organization. Hence, one should be cautious when extrapolating the results from the experiment to a potential scale-up by the government itself.[42]

The empirical results and the model have several policy implications. First, they show that interventions improving the speed of public service delivery are not necessarily effective tools for reducing corruption. Two important features of my setting are that there are no close substitutes to the public service and that the bureaucrats can control both the service delivery speed and whether the service is delivered at

[41] If instead, I use the effect on the reported payment, the total increase is US$2.1 million per year.

[42] If the scorecards were to be scaled up to all bureaucrats, there would also be a larger effect on the benchmark performance than what occurred in the experiment, where only half of the offices received the scorecards. This would shift the whole distribution of performance percentiles down. According to the model, this general equilibrium effect would induce more effort and smaller bribe payments than the partial experimental rollout of the scorecards.

all. Further research is needed to investigate if interventions targeting service delivery speed can also decrease corruption when bureaucrats can only control the speed of service delivery or when closer substitutes to the service are available.

Second, the differential effects of the scorecards on underperforming and overperforming offices suggest that it is especially important to improve information about underperforming bureaucrats. This implies that the type of recognition system that is common for bureaucrats—where outstanding performances are recognized without addressing inadequate performances—is ineffective. Positive feedback might still have an overall positive effect, but due to the reputation-level effect, it is less effective than negative feedback and can even be counterproductive.

Finally, the model points out a general problem when using performance feedback for socially desirable behavior. If the reputations or self-perceptions of some agents are improved, this may have negative spillovers on all other behavior where reputation is a motivating factor. When evaluating such interventions, it is therefore important to measure effects on all domains of performance, not just performances where there could be direct spillovers due to multitasking. This is an especially important insight for government bureaucracies, where compressed wage structures, secure employment, and potentially counterproductive incentives due to corruption often make reputational concerns and a sense of pride in one's work more important motivators.

REFERENCES

a2i, and Government of Bangladesh. 2020. "Data on Applications Made in e-governance System." Dhaka: Goverment of Bangladesh (last data provision December 2020).

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9–10): 1082–95.

Ashraf, Anik. 2022. "Performance Ranks, Conformity, and Cooperation: Evidence from a Sweater Factory." CESifo Working Paper 9591.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery." *Journal of Public Economics* 120: 1–17.

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee. 2014. "Awards Unbundled: Evidence from a Natural Field Experiment." *Journal of Economic Behavior and Organization* 100: 44–63.

Ayres, Ian, Sophie Raseman, and Alice Shih. 2013. "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage." *Journal of Law, Economics, and Organization* 29 (5): 992–1022.

Bai, Jie, Seema Jayachandran, Edmund J. Malesky, and Benjamin A. Olken. 2019. "Firm Growth and Corruption: Empirical Evidence from Vietnam." *Economic Journal* 129 (618): 651–77.

Banerjee, Abhijit V. 1997. "A Theory of Misgovernance." *Quarterly Journal of Economics* 112 (4): 1289–332.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2021. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy, and Training." *American Economic Journal: Economic Policy* 13 (1): 36–66.

Banerjee, Abhijit V., Esther Duflo, and Rachel Glennerster. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System." *Journal of the European Economic Association* 6 (2–3): 487–500.

Banerjee, Abhijit, Rema Hanna, and Sendhil Mullainathan. 2012. "Corruption." In *Handbook of Organizational Economics*, edited by Robert S. Gibbons and John Roberts, 1109–47. Princeton University Press.

Banuri, Sheheryar, and Philip Keefer. 2016. "Pro-social Motivation, Effort and the Call to Public Service." *European Economic Review* 83: 139–64.

Barrera-Osorio, Felipe, Kathryn Gonzalez, Francisco Lagos, and David J. Deming. 2020. "Providing Performance Information in Education: An Experimental Evaluation in Colombia." *Journal of Public Economics* 186: 104185.

BBS. 2014. "Bangladesh Population and Housing Census 2011." Bangladesh Bureau of Statistics (BBS). https://data.humdata.org/dataset/bangladesh-subnational-boundaries-and-tabular-data (accessed December 3, 2022).

BBS. 2023. "Bangladesh Subnational Administrative Boundaries." Bangladesh Bureau of Statistics (BBS). https://data.humdata.org/dataset/cod-ab-bgd (accessed January 1, 2023).

Bertrand, Marianne, Robin Burgess, Arunish Chawla, and Guo Xu. 2020. "The Glittering Prizes: Career Incentives and Bureaucrat Performance." *Review of Economic Studies* 87 (2): 626–55.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan. 2007. "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption." *Quarterly Journal of Economics* 122 (4): 1639–76.

Blader, Steven, Claudine Gartenberg, and Andrea Prat. 2020. "The Contingent Effect of Management Practices." *Review of Economic Studies* 87 (2): 721–49.

Byrne, David P., Andrea La Nauze, and Leslie A. Martin. 2018. "Tell Me Something I Don't Already Know: Informedness and the Impact of Information Programs." *Review of Economics and Statistics* 100 (3): 510–27.

Callen, Michael, Saad Gulzar, Ali Hasanain, Muhammad Yasir Khan, and Arman Rezaee. 2020. "Data and Policy Decisions: Experimental Evidence from Pakistan." *Journal of Development Economics* 146: 102523.

Clot, Sophie, Gilles Grolleau, and Lisette Ibanez. 2018. "Moral Self-Licencing and Social Dilemmas: An Experimental Analysis from a Taking Game in Madagascar." *Applied Economics* 50 (27): 2980–91.

Dal Bó, Ernesto, Frederico Finan, Nicholas Y. Li, and Laura Schechter. 2021. "Information Technology and Government Decentralization: Experimental Evidence from Paraguay." *Econometrica* 89 (2): 677–701.

de Janvry, Alain, Guojun He, Elisabeth Sadoulet, Shaoda Wang, and Qiong Zhang. 2022. "Subjective Performance Evaluation, Influence Activities, and Bureaucratic Work Behavior: Evidence from China." NBER Working Paper 30621.

Debnath, Sisir, Mrithyunjayan Nilayamgode, and Sheetal Sekhri. 2023. "Information Bypass: Using Low-Cost Technological Innovations to Curb Leakages in Welfare Programs." *Journal of Development Economics* 164: 103137.

Deserranno, Erika, Philipp Kastrau, and Gianmarco León-Ciliotta. 2022. "Promotions and Productivity: The Role of Meritocracy and Pay Progression in the Public Sector." STEG Working Paper 037.

Dhaliwal, Iqbal, and Rema Hanna. 2017. "The Devil Is in the Details: The Successes and Limitations of Bureaucratic Reform in India." *Journal of Development Economics* 124: 1–21.

Do, Quoc-Anh, Trang Van Nguyen, and Anh Tran. 2021. "Do People Pay Higher Bribes for Urgent Services? Evidence from Informal Payments to Doctors in Vietnam." Working Paper.

Dodge, Eric, Yusuf Neggers, Rohini Pande, and Charity Moore. 2021. "Updating the State: Information Acquisition Costs and Public Benefit Delivery." Working Paper.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–78.

Finan, Frederico, Benjamin A. Olken, and Rohini Pande. 2017. "The Personnel Economics of the Developing State." In *Handbook of Economic Field Experiments*, Vol. 2, edited by Abhijit Vinayak Banerjee and Esther Duflo, 467–514. Elsevier.

Freund, Caroline, Mary Hallward-Driemeier, and Bob Rijkers. 2016. "Deals and Delays: Firm-Level Evidence on Corruption and Policy Implementation Times." *World Bank Economic Review* 30 (2): 354–82.

Guriev, Sergei. 2004. "Red Tape and Corruption." *Journal of Development Economics* 73 (2): 489–504.

Hague Institute for Innovation of Law. 2018. *Justice Needs and Satisfaction in Bangladesh*. Hague Institute for Innovation of Law.

Hanna, Rema, and Shing-Yi Wang. 2017. "Dishonesty and Selection into Public Service: Evidence from India." *American Economic Journal: Economic Policy* 9 (3): 262–90.

Holmström, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7: 24–52.

Huntington, Samuel P. 1968. *Political Order in Changing Societies*. Yale University Press.

Kaufmann, Daniel, and Shang-Jin Wei. 1999. "Does 'Grease Money' Speed up the Wheels of Commerce?" NBER Working Paper 7093.

Khan, Adnan Q., Asim Ijaz Khwaja, and Benjamin A. Olken. 2019. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." *American Economic Review* 109 (1): 237–70.

Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken. 2016. "Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors." *Quarterly Journal of Economics* 131 (1): 219–71.

Khan, Muhammad Yasir. 2023. "Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan." Working Paper.

Leaver, Clare. 2009. "Bureaucratic Minimal Squawk Behavior: Theory and Evidence from Regulatory Agencies." *American Economic Review* 99 (3): 572–607.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–102.

Leff, Nathaniel H. 1964. "Economic Development through Bureaucratic Corruption." *American Behavioral Scientist* 8 (3): 8–14.

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review* 99 (1): 112–45.

Mattsson, Martin. 2021. *Performance Scorecards and Government Service Delivery: Experimental Evidence from Land Record Changes in Bangladesh*. AEA RCT Registry. https://doi.org/10.1257/rct.3232-3.2.

Mattsson, Martin. 2023. "When Does Corruption Cause Red Tape? Bribe Discrimination under Asymmetric Information." Preprint, SSRN. https://dx.doi.org/10.2139/ssrn.4395094.

Mattsson, Martin. 2024. *Data and Code for: "Information Systems, Service Delivery, and Corruption: Evidence From the Bangladesh Civil Service."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. http://doi.org/10.3886/E208342V1.

Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31 (4): 103–24.

Muralidharan, Karthik, Paul Niehaus, Sandip Sukhtankar, and Jeffrey Weaver. 2021. "Improving Last-Mile Service Delivery Using Phone-Based Monitoring." *American Economic Journal: Applied Economics* 13 (2): 52–82.

Niehaus, Paul, and Sandip Sukhtankar. 2013. "The Marginal Rate of Corruption in Public Programs: Evidence from India." *Journal of Public Economics* 104: 52–64.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–49.

Prendergast, Canice. 2007. "The Motivation and Bias of Bureaucrats." *American Economic Review* 97 (1): 180–96.

Raffler, Pia J. 2022. "Does Political Oversight of the Bureaucracy Increase Accountability? Field Experimental Evidence from a Dominant Party Regime." *American Political Science Review* 116 (4): 1443–59.

Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *Economic Journal* 128 (608): 413–46.

Reinikka, Ritva, and Jakob Svensson. 2005. "Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda." *Journal of the European Economic Association* 3 (2–3): 259–67.

Rose-Ackerman, Susan. 1978. *Corruption: A Study in Political Economy*. Academic Press.

Sachdeva, Sonya, Rumen Iliev, and Douglas L. Medin. 2009. "Sinning Saints and Saintly Sinners: The Paradox of Moral Self-Regulation." *Psychological Science* 20 (4): 523–28.

Svensson, Jakob. 2003. "Who Must Pay Bribes and How Much? Evidence from a Cross Section of Firms." *Quarterly Journal of Economics* 118 (1): 207–30.

Transparency International Bangladesh. 2015. Microdata for: *"Corruption in Service Sectors, National Household Survey 2015"* (accessed July 6, 2017).

Transparency International Bangladesh. 2016. *Corruption In Service Sectors, National Household Survey 2015*. Transparency International Bangladesh.

Transparency International Bangladesh. 2018. *Corruption in Service Sectors: National Household Survey 2017*. Transparency International Bangladesh.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *Quarterly Journal of Economics* 134 (2): 557–98.