Photographs: © UNICEF - Pakistan

Command and Can't Control: An Evaluation of Centralized Accountability in the Public Sector*

DeliverEd Initiative Working Paper

Saad Gulzar, Juan Felipe Ladino,
Muhammad Zia Mehmood, and Daniel Rogger

March 2023

the Education Commission
UKaid from the British people
Blavatnik School of Government
University of Oxford
Bureaucracy Lab
Institute of Development & Economic Alternatives

DeliverEd: Building knowledge on how to use delivery approaches to advance education reforms

The DeliverEd Initiative was launched in 2019 to strengthen the evidence base for how governments can achieve their policy priorities through delivery units and other delivery approaches. Globally, more than 80 countries have used such approaches to achieve better outcomes for policy reform and implementation. Forty-seven percent of those include an education focus, either as a single focus sector or as part of a multisector approachi. But there was little empirical evidence, especially from developing countries, on the effectiveness of delivery approaches in delivering education outcomes or on the design choices, contextual features, and enabling factors that contribute to their performance.

DeliverEd has helped to fill this evidence gap and create a better understanding of the practices leaders can adopt to improve their policy delivery and reform efforts. It has conducted research within and across countries on the effectiveness of delivery approaches in improving reform implementation, with the key findings included in this final report. It has facilitated knowledge and experience sharing among countries—for example, through the Africa Policy Forum—to equip policymakers with a deeper understanding of delivery challenges and solutions to make informed decisions. It continues to increase awareness and the uptake of research to improve schooling and learning in low-income countries.

The Education Commission leads DeliverEd with Oxford University's Blavatnik School of Government and funding from the UK Foreign, Commonwealth, and Development Office (FCDO). Other partners include the University of Toronto, the Institute for Educational Planning and Administration (under the Auspices of UNESCO), University of Cape Coast, Ghana, Institute of Development and Economic Alternatives (IDEAS) in Pakistan, World Bank, and Georgetown University in the U.S. For more information about DeliverEd, and to view the country studies and other related research and policy engagement materials, please visit www.educationcommission.org/delivered-initiative.

We are very grateful to the Blavatnik School of Government and all our research partners for their in-depth research, especially during the COVID-19 pandemic. This DeliverEd Final report is the Education Commission's interpretation of the research. For the detailed research papers themselves, please see the next page.

DeliverEd Research Products

Anderson, K., Ibarra, A., & Javaid, N. (December 2022). The Education Commission. A Case Study of the Sierra Leone Delivery Unit. DeliverEd Initiative Policy Note.

Bell, S., Asim, M., Mundy, K., Pius Nuzdor, H., Boakye-Yiadom, M., & Mensah Adosi, C. (May 2023). How do regions, districts and schools respond to the introduction of a delivery approach: Evidence from Ghana. DeliverEd Initiative Working Paper.

Bell, S., Leaver, C., Mansoor, Z, Mundy, K., Qarout, D., & Williams, M. (March 2023). The Role of Delivery Approaches in Education Systems Reform: Evidence from a Multi-Country Study. DeliverEd Initiative Working Paper.

Boakye-Yiadom, M., Leaver C., Mansoor, Z., & Iocco, MP. (March 2023). Management and performance in mid-level bureaucracies: Evidence from Ghanaian education districts. DeliverEd Initiative Working Paper.

Gulzar, S., Ladino, JF., Mehmood, MZ., & Rogger, D. (March 2023). Command and Can't Control: An Evaluation of Centralized Accountability in the Public Sector. DeliverEd Initiative Working Paper.

Malik, R. & Bari, F. (May 2023). Improving Service Delivery via Top-Down Data-Driven Accountability: Reform Enactment of the Education Road Map in Pakistan. DeliverEd Initiative Working Paper.

Mansoor, Z., Qarout, D., Anderson, K., Carano, C., Yecalo-Tecle, L., Dvorakova, V., & Williams, M. (July 2021). A Global Mapping of Delivery Approaches. DeliverEd Initiative Working Paper.

Qarout, D. (November 2022). The Accountability Paradox: Delivery Units in Jordan's Education Sector 2010–2019. DeliverEd Initiative Working Paper.

Williams, M., Leaver, C., Mundy, K., Mansoor, Z., Qarout, D., Asim, M., Bell, S., & Bilous, A. (April 2021). Delivery Approaches to Improving Policy Implementation: A Conceptual Framework. DeliverEd Initiative Working Paper.

DeliverEd Policy Products

Anderson, K., & Bergmann, J. (2022). The Education Commission. Design Choices for Delivery Approaches in Education, DeliverEd Initiative Policy Brief.

Anderson, K., & Carano, C. (2021). The Education Commission. The Challenge of Delivering for Learning DeliverEd Initiative Policy Brief.

Villaseñor, P. The Education Commission. (2021) Delivery Approaches in Crisis or Conflict Situations, DeliverEd Initiative Policy Brief.

Abstract

By reducing information asymmetries across the hierarchy, the digitization of government services presents an opportunity for centralized management of frontline staff. In particular, high-frequency granular data can enable senior government officials to hold poorly performing members of the service delivery chain to account. To be effective, however, centralized management must translate large volumes of data into appropriate management actions. This paper studies this tension by evaluating a large-scale centralized accountability approach to managing education carried out at scale in Punjab, Pakistan. We find that a system that automatically identified poorly performing schools and jurisdictions for the attention of central management had no appreciable impact on the trajectory of school outcomes across any area of its focus. We contrast this result with the significant impact that frontline managers (head teachers) can have on school outcomes across the same areas and the potential for using centralized information systems to optimize the allocation of managerial talent across the public sector.

JEL CODES: D73, H11, H83

[*] We gratefully acknowledge funding from the Blavatnik School of Government/Education Commission DeliverEd program, and the World Bank's i2i initiative, Knowledge Change Program, and Governance Global Practice. We thank Belen Torino for excellent research assistance and our counterparts at the Punjab Program Monitoring and Implementation Unit for providing us with data and details of the institutional environment. Finally, we thank Faisal Bari, Clare Leaver, Rabea Malik, and Martin Williams for their guidance. All errors are our own. The views expressed in this chapter are those of the authors alone, and should not be attributed to the World Bank, its executive directors, or the countries they represent.

†Gulzar: Department of Politics and School of Public and International Affairs, Princeton University;
Ladino: Department of Economics, Stockholm University; Mehmood: Haas Business School, University of California Berkeley; Rogger: World Bank Development Impact Evaluation Research Department.

1. Introduction

A fundamental policy question is how to improve the management of the state toward pro- viding high-quality public services. Detailed monitoring of public sector actors has been proposed as a strategy to promote improved performance and better development outcomes (Finan, Olken, & Pande, 2015; Duflo, Hanna, & Ryan, 2012). However, improving public services through greater oversight must overcome classical measurement and incentive prob- lems inherent to the public sector. This paper investigates the efficacy of such an approach by evaluating the effectiveness of a showcased centralized public sector monitoring system.

The digitization of government services, and the modern capacity to generate data, has ex- panded the possibility managers have for oversight of the public sector. By reducing the cost of observing measures of service delivery and the activities of public sector actors and organizations, digitization has reduced information asymmetries across the hierarchy. In particular, greater access to a variety of measures of the functioning of government allows public sector managers to better distinguish poor performance and to hold poorly performing members of the service delivery chain to account. This has the potential to flatten govern- ment hierarchies and create a closer link between senior officials and the activities for which they are held accountable by citizens.

However, the finite dimensions of any management information system means that man- agement actions that attempt to address fast-moving and context-dependant outcomes may struggle to identify appropriate responses. The multi-faceted nature of many routine public sector activities may inhibit centralized, data-based, management of government. Without proximity to the wider context in which measures of government functioning are evolving, centralized management may be unable to identify appropriate management actions or ac- curately identify their effects.

Schemes that emphasize centralized oversight must also overcome constraints imposed by the sheltered incentive environment of many governments (Besley, Burgess, Khan, & Xu, 2022). Unable to substantially shift de jure contracting environments, senior managers of the public service may not be able to motivate work teams to solve routine challenges. As such, centralized delivery approaches have typically emphasized the use of de facto authority to generate changes in the performance of frontline staff. For example, senior managers may seek to threaten reputational damage or shaming in front of public sector peers. Severe cases may lead to reposting of personnel to undesirable positions, an element of the de facto public service contract that has been shown to be highly motivating.

In this paper, we explore how senior government officials' high-frequency monitoring of public services impacts subsequent performance. We focus on a centralized government monitoring system in the education sector of Punjab, the largest province in Pakistan, that generates monthly performance reports for senior managers across a range of key measures of service quality. Specifically, we study the impacts of flagging poor performance on teacher presence, student attendance, functional facilities, and students' test scores on standardized exams.

The scheme we study is a showpiece of centralized delivery models. It is advised by individ- uals who are considered the top experts in the field, with the full backing and involvement of the most senior members of the executive, and it has been implemented to a very high standard for over six years (Barber, 2013; Chaudhry & Tajwar, 2021; Malik & Bari, 2022). Reviewing the scheme in an interview with the Government of Punjab in 2017, Michael Barber stated, "Punjab is unique ... across the whole world for combining deliverology with really good and modern technology." Our empirical assessment supports the assertion that the scheme was implemented as intended to a high degree of fidelity.

In the Punjab scheme, monthly data was collected on the performance of all 52,000 schools that existed in the province between December 2011 and May 2018. The process resulted in a total of nearly two million observations of school quality across the period. A reporting system presented managers throughout the Punjab Government with frequently updated information on the state of schools within their area of responsibility and in the province as a whole. We provide evidence that the data was accurately collected and the reports were disseminated as planned.

The chief minister, the highest member of the executive, personally chaired regular "stock- takes" of the education sector based on this data. Throwing his own personal reputation and political capital behind the scheme, he created a climate of de facto accountability for service failures. Remarks by the chief minister of Punjab at the time indicate how much political capital he invested in the scheme. A qualitative review of the scheme stated that "At the core of the approach design was leveraging political interest and political capital to orient the bureaucratic structures involved in service delivery toward improvements at a fast pace." (Malik & Bari, 2022). The intention was that with strong support from senior management, and de facto threats of punishment, officials throughout the system would respond to any managerial directives that aimed to address areas of poor performance.

Overall, the scheme we study was designed by renowned experts in the field, had the weight of the most senior members of government behind it, and was executed effectively. As such, it is a useful setting to evaluate centralized management of frontline services. Using a stacked difference-in-differences design suggested by Cengiz, Dube, Lindner, and Zipperer (2019) and Baker, Larcker, and Wang (2022), we assess the impact of being flagged in the monitoring system as severely underperforming on a number of key educational outcomes.

Our design compares schools in administrative units that have been flagged with schools with similar observables (and in some specifications the same history of previous flagging) in administrative units that have not been flagged, both of which see a comparable drop in the variable of interest. A school is only flagged if a sufficient number of schools within the same district also have fallen in the variable of interest. This allows us to identify schools with similar characteristics but in areas where the performance of other schools fell in such a way in a particular month that the administrative unit ends up with an average level of the variable of interest just above and below the flagging threshold.

We find no evidence that the scheme had significant impacts on school quality across the range of measures that it targeted. Rather, we show that the flagging system correctly captures negative transitory shocks in schools. However, those schools in the administrative areas that get flagged follow a very similar pattern of return to their equilibrium state of service delivery as their comparison schools in administrative areas that were not flagged. The transitions of both sets of schools appear to follow a reversion to the mean. We are also able to assess whether flagging creates any significant policy action from local bureaucrats. We find no evidence of more visits from these bureaucrats to affected schools, financial investments, or transfers of teachers and head teachers.

As such, it appears that the elements of public sector management implemented in Punjab's education sector that focused on data-driven oversight and accountability did not have ap- preciable impacts on education service delivery or outcomes. Though there were significant investments by centralized actors in designing and implementing the system to what was perceived as a high standard, it did not have the intended impacts.

A key question is, therefore, why the scheme continued for so long and has received plaudits from the centralized actors involved. Barber (2021) states that “once the trickle of early progress became a strong current, the sense of momentum carried them . . . what had once seemed impossible became routine.” One possibility is that without defining an appropriate counterfactual, the scheme was judged by assessing the immediate impacts for service delivery of flagging. Using a naive estimator of the impacts of flagging, we observe a positive and significant impact, which may have been interpreted as a positive effect of the monitoring scheme. Once this idea was in place, observational evidence for it could be found everywhere.

However, the absence of a rigorous understanding of the scheme’s impacts by senior managers points to an important point about the use of large-scale data in the public sector. With our evaluation approach, we estimate that the null effects of the scheme could have been detected within months of implementation. The impact of large-scale data collection on public sector effectiveness is mediated by the existence of an analytical team with an understanding of counterfactual inference. Such a team was absent from Punjab, but it could have quickly identified the true effect of the scheme and helped to reform it or make clear the limitations of centralized management of frontline service provision.

Moreover, such a team could have capitalized on the power of ‘coarse’ but large-scale data analysis to identify structural parameters in the delivery system. We present one such set of parameters by estimating the relative talents of individual frontline managers -head teachers-. As a benchmark for the null effects of the centralized management scheme, we highlight that the impact of head teacher management can be substantial. Using the same monitoring data, we show that a program that focused on reallocating head teachers based on matching their talents and school needs could have substantially improved educational outcomes.

As such, our paper contributes to the literature on multiple areas of public sector management. The paper is closely related to the study of the effects of top-down monitoring on the performance of bureaucrats and its efficacy relative to empowering more autonomous public sector agents. Evidence on the impact of monitoring on performance is mixed, with generally though not universally positive results from frontline settings (Olken, 2007; Hussain, 2015; Dhaliwal & Hanna, 2017; Callen, Gulzar, Hasanain, Khan, & Rezaee, 2020; Duflo et al., 2012; Muralidharan, Niehaus, Sukhtankar, & Weaver, 2021); and less supportive evidence from experiments that investigate the impacts on individual motivation and performance (Falk & Kosfeld, 2006; Dickinson & Villeval, 2008; Bandiera, Best, Khan, & Prat, 2021). We extend this literature by providing evidence on the impacts of centralized oversight to the broader administrative environment. Classical theories of monitoring in public administra- tion (Dixit, 2002) and existing descriptive studies (Rasul & Rogger, 2018; Rasul, Rogger, & Williams, 2020) imply that monitoring may be less successful in such administrative settings.

Moreover, our study adds to the nascent literature on designing optimal management struc- tures in the public sector (A. Banerjee, Chattopadhyay, Duflo, Keniston, & Singh, 2021; Ali, Fuenzalida, G´omez, & Williams, 2021). Muralidharan and Singh (2020) show that a large- scale program to improve managerial practices in India's education administration did not improve school-level outcomes or student scores. We benchmark our results on centralized oversight by showcasing how central management might use large-scale measures of school outcomes to support empowered but autonomous frontline managers in having maximum impacts on service delivery. We thus add to the evidence base for drivers of performance in settings that suffer from constraints on the use of punishments in the public sector (B´ekir, Harbi, Grolleau, Mzoughi, & Sutan, 2016; Belot & Schr¨oder, 2016).

Our study evaluates a celebrated and at-scale use of a monitoring approach to managing public service delivery. It therefore adds to the literature on understanding program impacts for schemes implemented by government, argued to be an important test of the external validity of pilot programs (Muralidharan & Niehaus, 2017; Vivalt, 2020). First, the richness of our data allows us to explore a wide range of potential behavior by bureaucratic actors across the service delivery chain in response to the scheme. Second, the paper provides a lens through which to understand results of smaller pilots of centralized oversight, such as Callen et al. (2020), who show that flagging underperforming health facilities in Punjab positively affected subsequent public health workers' attendance. When taken to scale, such pilots may not provide a sustainable means of managing the public administration (A. V. Banerjee, Duflo, & Glennerster, 2008; A. Banerjee et al., 2021).

This paper also contributes to a more general strand of the literature that explores the effects of incentives on bureaucracies in developing countries (Finan et al., 2015). Such incentives include financial rewards (Muralidharan & Sundararaman, 2011; Dal B´o, Finan, & Rossi, 2013; Ashraf, Bandiera, & Jack, 2014; Deserranno, 2019; Leaver, Ozier, Serneels, & Zeitlin, 2021), career incentives (Khan, Khwaja, & Olken, 2019; Bertrand, Burgess, Chawla, & Xu, 2020; Deserranno, Leon, & Kastrau, 2022), or other non-financial incentives (Ash & MacLeod, 2015; Honig, 2021). The nature of the scheme we evaluate indicates that even with substantial political backing, there may be limits to the extent that punishments can be systematically exploited in a public administration setting to motivate sustained change.

Finally, our paper contributes to the large existing literature on determinants of educational outcomes in low- and medium-income countries (Center for Global Development, 2022; World Bank Group, 2020; Glewwe & Kremer, 2006). More specifically, we add to existing evidence that speaks to the value of effective teachers (Bau & Das, 2020; Crawfurd & Rolleston, 2020; Chetty, Friedman, & Rockoff, 2014; Hanushek & Rivkin, 2006; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004) and head teachers (Coelli & Green, 2012; Liebowitz & Porter, 2019; Branch, Hanushek, & Rivkin, 2012), and the emerging evidence on the management of schools and school districts (Mbiti, 2016; Lemos, Muralidharan, & Scur, 2021; Leaver, Lemos, & Scur, 2019).

The paper proceeds as follows: Section 2 describes the setting and structure of the public service we study. Section 3 describes the centralized monitoring scheme. Section 4 introduces the data and presents our empirical approach. Section 5 describes and discusses our evaluation of the high-frequency monitoring scheme for multiple tiers of government. Section 6 discusses the reasons for the persistence of the high-frequency monitoring scheme. Section 7 illustrates an alternative use of the monitoring data that focuses on determining the quality of frontline managers (head teachers) and using that information to improve educational outcomes. Finally, Section 8 concludes.

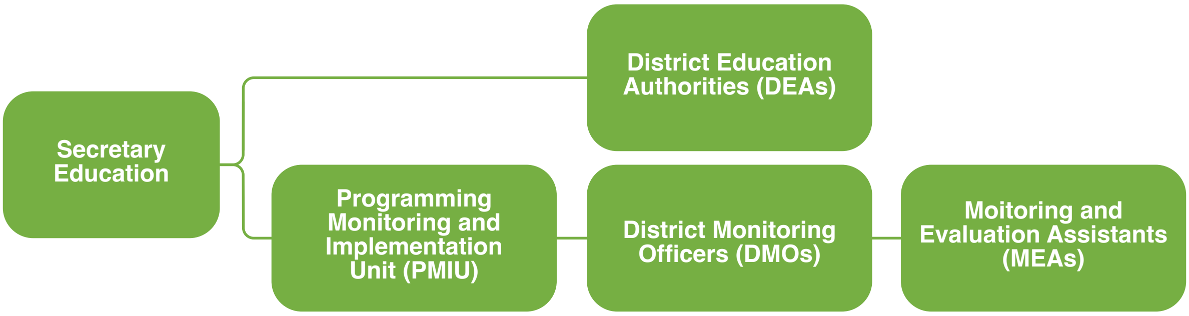
2. Public Education in Punjab

Punjab is Pakistan's most populous province and is home to over 110 million people, half of the country's population. Twenty million people are school-aged children, many of whom attend school, of which there are approximately 52,000, with 400,000 teachers (School Edu- cation Department, 2018). The scale of managing education in the province is substantial.

The province is divided into 36 districts, which are further subdivided into administrative sub-units called tehsils, which are further subdivided into areas of responsibility called "maraakiz."[1] There are an average of four tehsils per district and an average of 48 maraakiz per tehsil. Thus, any district education manager has on average 192 administrative units, with associated staff and schools, to track. Even at the most granular administrative level, a markaz official must manage an average of 20 schools.

The School Education Department is responsible for organizing and overseeing the perfor- mance of the education sector. As Figure 1 describes, the department has two arms: district education authorities (DEAs), which coordinate the implementation of public education de- livery, and the Program Monitoring and Implementation Unit (PMIU), which is responsible for independently collecting and disseminating data on performance of schools within the department. The two arms of the education administration are staffed and organized sepa- rately, and monitoring is generally seen as independent of implementation.

Figure 1: School Education Department of Punjab


Secretary Education
District Education Authorities (DEAs)
Programming Monitoring and Implementation Unit (PMIU)
District Monitoring Officers (DMOs)
Moitoring and Evaluation Assistants (MEAs)

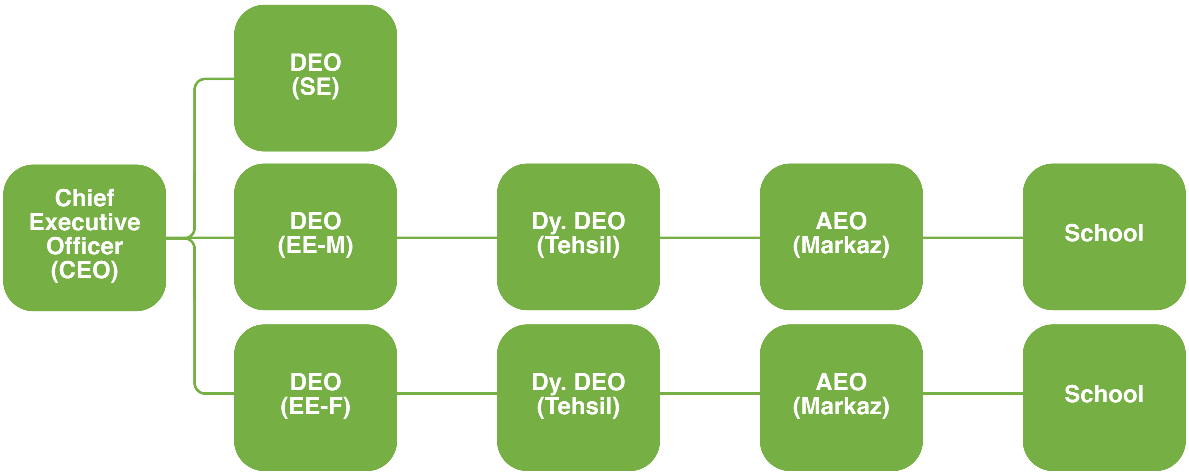
2.1 District education authorities

There are 36 DEAs, one for each district, led by an education chief executive officer (CEO), who reports to the provincial education secretary. Below the CEO are layers

[1] Plural of the term "markaz," the Urdu word for "center."

of hierarchy that correspond to the district's administrative organization: district education officers (DEOs), two deputy district education officers (DDEOs) for each tehsil, and assistant education of- ficers (AEOs) at the markaz level. Figure 2 illustrates the reporting structure within each of the 36 DEAs of Punjab. Schools in the district are categorized into one of three wings: elementary education female (EE-F), elementary education male (EE-M), and secondary education (SE). Our study focuses on the EE-M and EE-F wings, which comprise primary schools (children aged approximately 4 to 9) and middle schools (children aged approximately 10 to 12). Together, these schools make up roughly 80 percent of all public schools in the province.

Figure 2: District Education Authorities



All layers of the hierarchy are expected to manage the work of those officers under them and are able to censure underperforming school staff. AEOs can be seen as a layer of hierarchy above school principals and thus complete the chain of command from senior management to the school level.

Such a layered hierarchy is not unusual in administrative settings of this scale across the world. The simple physical constraint of traveling to schools, engaging with head teachers and handling administrative tasks associated with those schools implies a limit on the scale of any individual official's ability for oversight.

By contrast, a feature of large-scale measurement and aggregation in management informa- tion systems is that it can alleviate these physical constraints and centralize the ability to supervise and censure at scale. By dramatically lowering the cost of monitoring individual schools and jurisdictions, digitization has opened up the possibility of centralized manage- ment throughout the hierarchy.

2.2 The PMIU

While the DEAs are responsible for outcomes in public schools at the district level, the Program Monitoring and Implementation Unit (PMIU) is tasked with monitoring the per- formance of the DEAs. The PMIU conducts both an annual census of all public schools in the province, as well as a monthly monitoring of those schools to assess key aspects of the school environment from the quality of infrastructure to teacher presence. Undertaking these duties are monitoring and evaluation assistants (MEAs), typically retired army officers hired just for collecting data and not answerable to DEAs.

Across the period we study, the MEAs collected performance-related data from every school on an unannounced random date of every month. The assignment of monthly school inspec- tions to MEAs was randomized to limit collusion between the MEAs and school staff. As will be discussed below, our analysis of the consistency of different sources of data on schools implies that this process produced valid assessments of school functioning.

Data collected by the PMIU was the basis of performance reports generated every month and called "datapacks." These datapacks were the center of monthly and quarterly meetings of senior education managers in the province. They were first generated in December 2011 and then prepared monthly during the school year from then on. We study the period up until May 2018, just before the national elections and a change in administration. The datapacks reported for each district the aggregate performance at the markaz level along multiple dimensions: teacher presence, student attendance, visits by DEA staff, and status of school facilities (electricity, drinking water, toilets, and the boundary wall).2 From September 2017, the datapacks also reported scores on standardized math, English, and Urdu tests. The MEAs administered tests in these subjects to seven randomly selected grade three students from each school they visited as part of their monthly inspections.

The reported performance on each dimension of school functioning was color-coded in the datapacks based on standardized performance thresholds set by the chief minister's team. A markaz could be coded red, orange, or green, with red being the primary flag for underper- formance. Figure A1 in the Appendix illustrates the color-coding in part of the April 2013 datapack for the district of Rajanpur. As such, a

[2] The datapacks also included the number of schools surveyed for each markaz, whether they were found closed at the inspection visits, a breakdown of statistics by male and female schools, and recommendations for the markaz about which schools to focus on to improve aggregate outcomes.

specific AEO would be associated with any flagging of underperformance, and above her/him a DDEO and DEO. Flagging was also undertaken on tehsils and districts, but in a less systematic way. The focus of the discussions was on markaz performance, and so that is the emphasis we follow in our empirical work.

3. Centralized oversight intervention

In the context of the existing administration, and armed with the PMIU-generated data on school performance, the chief minister of Punjab set up a centralized oversight regime for education in 2011. He personally chaired the oversight committee and worked closely with the McKinsey International consultancy firm and a high-level advisor from the UK with expertise in centralized delivery approaches.

Figure 3 describes the structure of the monitoring regime. In month t data is collected on all schools in the province, and markaz-level averages (means) on performance are presented to senior managers in month t + 1. The maraakiz that do not reach specific (standardized) thresholds are flagged red or orange. These reports were the basis of senior management check-ins within the first 10 days of every month. Quarterly meetings were held where “the [chief minister] at that time was himself very very motivated and he would make it a point to not miss any one of the meetings.”[3] The senior management of the province placed a substantial weight on the system. The chief minister “had full ownership of this reform and [sent] a signal to the bureaucracy that they were to take it seriously” (Malik & Bari, 2022, p. 22).

There is no evidence that senior managers changed de jure contracting, such as making salaries conditional on performance. In a number of cases, there is evidence that one-off financial bonuses were given to the most successful districts’ CEOs, but not to AEOs or mid-level bureaucrats. In our empirical work, we explore whether there is evidence of staff transfers or long-term impacts on career trajectories resulting from poor performance, and we do not find any such evidence. Rather, it seems that senior management was constrained by public service rules meant to protect officials from political influence and instead had to rely on de facto incentives to punish officials they perceived as underperforming.

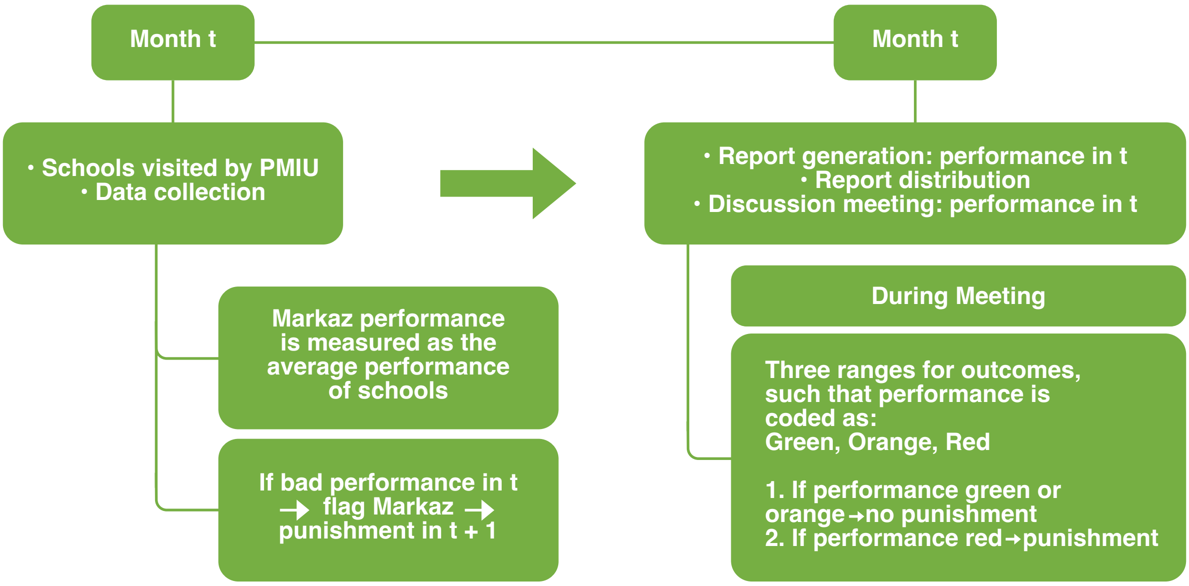
Interviews with department officials revealed that these meetings mostly involved the

[3] Referring to the quarterly oversight meetings, (Malik & Bari, 2022) state that “All other practices of priority setting, target setting and use of data for monitoring were all feeding into the construction of this accountability mechanism that was arguably central to the design of the delivery approach that was instituted in Punjab.”

officers flagged red in the datapack getting censured in front of their peers. As a district official characterized it, "the red were reprimanded and the greens were appreciated." Another dis- trict officer stated, "The constant monitoring by the Chief Minister and the Chief Secretary played a very critical role." Another said, "We do not want to be punished in front of our colleagues." As the chief minister's staff officer recounts, "I wouldn't say it was fear necessar- ily but the point [is] that the quarterly rankings and the performance accountability caused a lot of concern" (all quotations from Malik and Bari (2022)).

Concerned with the potential for being censured at meetings with the chief minister, dis- trict officials had incentives to motivate their subordinates, and their subordinates those below them, and so on. The intention of the scheme was that greater oversight by senior management would allow targeting of sanctions toward parties that most required motivation through the chain of command. As such, the scheme relied on the interaction between measurable outcomes and personnel management throughout the hierarchy. In public sector oversight models, while the output space can be collapsed to observable quantities, im- provements in these outputs still rely on multidimensional and non-contractible personnel management activities. As such, the question under evaluation is whether oversight and accountability regimes are an effective means of motivating more effective personnel

Figure 3: Monitoring Scheme Structure


Month t
Month t
• Schools visited by PMIU
• Data collection
• Report generation: performance in t
• Report distribution
• Discussion meeting: performance in t
Markaz performance is measured as the average performance of schools
If bad performance in t → flag Markaz → punishment in t + 1
During Meeting
Three ranges for outcomes, such that performance is coded as: Green, Orange, Red
1. If performance green or orange→no punishment
2. If performance red→punishment

[4] Datapacks were not produced for months of the calendar year in which schools were on summer break: June, July, and August.

man- agement throughout the hierarchy.

The political weight and international guidance behind the scheme ensured that it was effectively implemented. Datapacks were indeed produced for the 60 months from December 2011 to May 2018 as intended.4 As a measure of the quality of the data contained in the datapacks, we compared it to the Annual Census of Schools for the month in which the annual census was collected. Both data sources reported information about the number of teachers posted, enrolled students, and the functionality of school infrastructure. Figure B2 in the Appendix compares the sources. Panels (a) and (b) plot the distribution of the log-transformed teacher presence and student attendance from both sources from 2012 to 2017, respectively, as well as the distribution of errors when comparisons are for reports by the two data sets regarding the same school. Both in the overall distribution and school by school, it can be observed that the overall level of error is low. Similarly, Panel (c) plots the percentage of schools where the report of functionality in different infrastructures coincides: in both sources, the school reported the infrastructure as functional or not functional. The figure suggests that for all the years, the percentage of coincidence is near 100 percent.

Though there were slight modifications to the structure of the scheme, these elements re- mained at its core. The design is a demonstration case of centralized, data-informed ac- countability regimes. The centrality of the scheme to the administration's management, the scale and quality of data collection, and the length of time that the scheme was in place all make the scheme a good test for the efficacy of such approaches in the public sector.

4. Evaluation methodology

4.1 Data

We used monthly administrative data collected at the school level by the PMIU's MEAs from December 2011 to May 2018. The data excludes June, July, and August of each year, corresponding to summer vacations and public schools being closed. The primary outcomes included in the data set are monthly assessments of teacher presence, student attendance, and the extent to which the facilities are functional. Teacher presence and student attendance are measured as the percentage of teachers/students present at the time of the visit by the MEA. The functional facilities measure focuses on the status of four school facilities: drinking water, electricity, toilets, and the boundary wall. Additionally, starting in September 2017, PMIU began collecting data on student test scores in math, English, and Urdu using standardized tests administered by MEAs during their monthly school visits to seven randomly selected grade three students in each school. Scores are measured as the percentage of correct answers in the tests. To understand the effect on bureaucratic behavior, we also use the data collected on DEA visits to schools. For each school, we can identify its district, tehsil and markaz, and the history of flagging across administrative tiers and units.

Schools in this setting are relatively small, with an average of 4.6 teachers and 110 students. Roughly three percent of the schools have ever had more than 20 teachers, and these are evenly spread across the province. At the markaz level, there is a substantial variation in the number of schools within a markaz, broadly following differences in population size. However, the average number of schools that an AEO must manage is 20, of which close to 80 percent are elementary schools.

Over the entire period studied, 82 percent of maraakiz were flagged red at least once on some outcome measure, and 96 percent were flagged red or orange. Like any population of schools, there were some which were persistently high performers. Of the schools, 1.6 percent never dropped below 90 percent on any of the outcome measures. However, of the 82 percent of maraakiz flagged once, 79 percent got flagged again at some point. Thus, the oversight intervention was broad in its reach across maraakiz.

Panel A in Table 1 shows descriptive statistics for outcomes at the outcome-school-month level, differentiating between school outcomes located in maraakiz that were flagged on that outcome in a particular month and those in maraakiz that were not flagged on that outcome in that month. By construction, for all outcomes the mean of schools in a flagged markaz is lower than those in a non-flagged markaz. An equivalent set of descriptives is provided in Panel B for the markaz level, with the unit of observation the outcome-markaz-month. Both panels show that in the month in which a markaz is flagged on a particular outcome, there is a clear drop in the mean level of that outcome. Comparison of the two sets of columns provides a sense of the order of magnitude of change the scheme was aiming to engender. The flagged maraakiz have an average teacher presence of 80 percent

[5] Since this activity was based on a ranking, even if all districts were systematically improving, the ranking system kept rewarding districts with the highest relative scores and punishing those with the lowest scores.

while the non-flagged maraakiz have an average teacher presence of 93 percent.

Though flagging was widespread, there was a high degree of persistence in overall perfor- mance across schools. At the district level, poorly performing districts stayed poorly perform- ing throughout the period we studied. In addition to monthly flagging of AEOs/maraakiz, the districts were ranked each quarter. The ranking was defined for each meeting based on an overall score that took into account the performance of all the variables in the previous months.5 Panel C in Table 1 shows descriptive statistics for the overall scores on which the districts are ranked for those districts in the top or bottom positions. Once again by construction, bottom districts report a significantly lower mean of the overall score than the top ones. Panel C shows the percentage of districts that entered the top/bottom five positions in each period. Note that for all the districts on all the meeting dates, there is a relatively small number of cases where new districts fell into the top (7.7 percent) or bottom (8.3 percent) positions, which suggests that at the district level, there existed a high degree of persistence in the ranking status of those that were highlighted in the oversight meetings.

Figure 4 presents this persistence graphically. For each date of quarterly meetings, we color- coded the quintile in which the district fell in the distribution of the overall score. Although there was variation, the districts located in the higher quintiles of the distribution tended to maintain their high position in the ranking, while the districts in the lowest quintiles of the scores distribution remained in the lowest positions. The figure thus presents a descriptive sense that the flagging did not motivate poor performers sufficiently for their overall rankings to change.

How do we reconcile this feature of the intervention environment: almost all maraakiz were flagged at some point, and yet some districts and maraakiz were systematically at the bottom of the distribution? Evidence from other settings indicates that education (and other service delivery environments) face structural underlying constraints to improving outcomes (World Bank Group, 2018). However, they are also buffeted by a range of shocks (such as teachers getting sick) that substantially shift around the absolute levels of service delivery. This would imply that Punjab's schools face shocks that sometimes push them under the flagging threshold irrespective of their baseline levels of performance.

The time series variation in outcomes among the schools we studied is consistent with this interpretation. Table 2 presents the standard deviations in outcomes for schools in each quintile of mean baseline performance. The top four quintiles of schools face comparable levels of variation. There is some meaningful probability of falling below the thresholds in each. This probability is almost a magnitude higher in the lowest/first quintile. The probability of flagging jumps toward the bottom of the distribution, implying a persistently challenging environment to manage.

4.2 Empirical strategy

To estimate the effect of the high-frequency monitoring system on educational outcomes, we followed Cengiz et al. (2019) and Baker et al. (2022) to build a stacked data set to avoid biases driven by the time-varying nature of the treatment (De Chaisemartin & d'Haultfoeuille, 2020; Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021). The stacking consists in creating event-specific data sets for which we can identify a set of control units that have not been treated during a specific period. The process is described in Figure B1. The result is a data set with the treatment centered in relative time to eliminate its time-varying nature, conditional on indexing the estimations at the event-panel level. Following the stacked design of our data, we implemented a stacked difference-in-differences strategy.

4.2.1 Markaz flagging

Our main specifications assess the impact of a markaz being flagged as red/underperforming on the flagged outcomes in schools in that markaz. We estimated the following equation:

$$\begin{aligned}Y_{smdte} = \gamma_1(T_{mde} \times Flag_{te}) + \gamma_2(T_{mde} \times Punish_{te}) +\\ \beta(T_{mde} \times AfterFlag_{te}) + \alpha_{mde} + \lambda_{te} + dt + \epsilon_{smdte}\end{aligned}$$

Where subscripts s, m, d, t, e account for school, markaz, district, time, and event panel, respectively. All of the equation components indexed at the event panel e. Ysmdte represent any of the outcomes for school s, within markaz m, in district d. Tmde equals 1 for schools in a flagged markaz m. Flagte equals 1 for the period during which data is collected and the flag is defined, and Punishte equals 1 for the period just after the flagging where the oversight committee meets and the accountability intervention occurs. AfterFlagte equals 1 for the periods after the punishment phase where we assess intervention impact. αme is for markaz fixed effects to control for constant characteristics of maraakiz, and λte is for time fixed effects to capture time-specific shocks. We include dt—a district binary and linear

Table 1: Descriptive Statistics

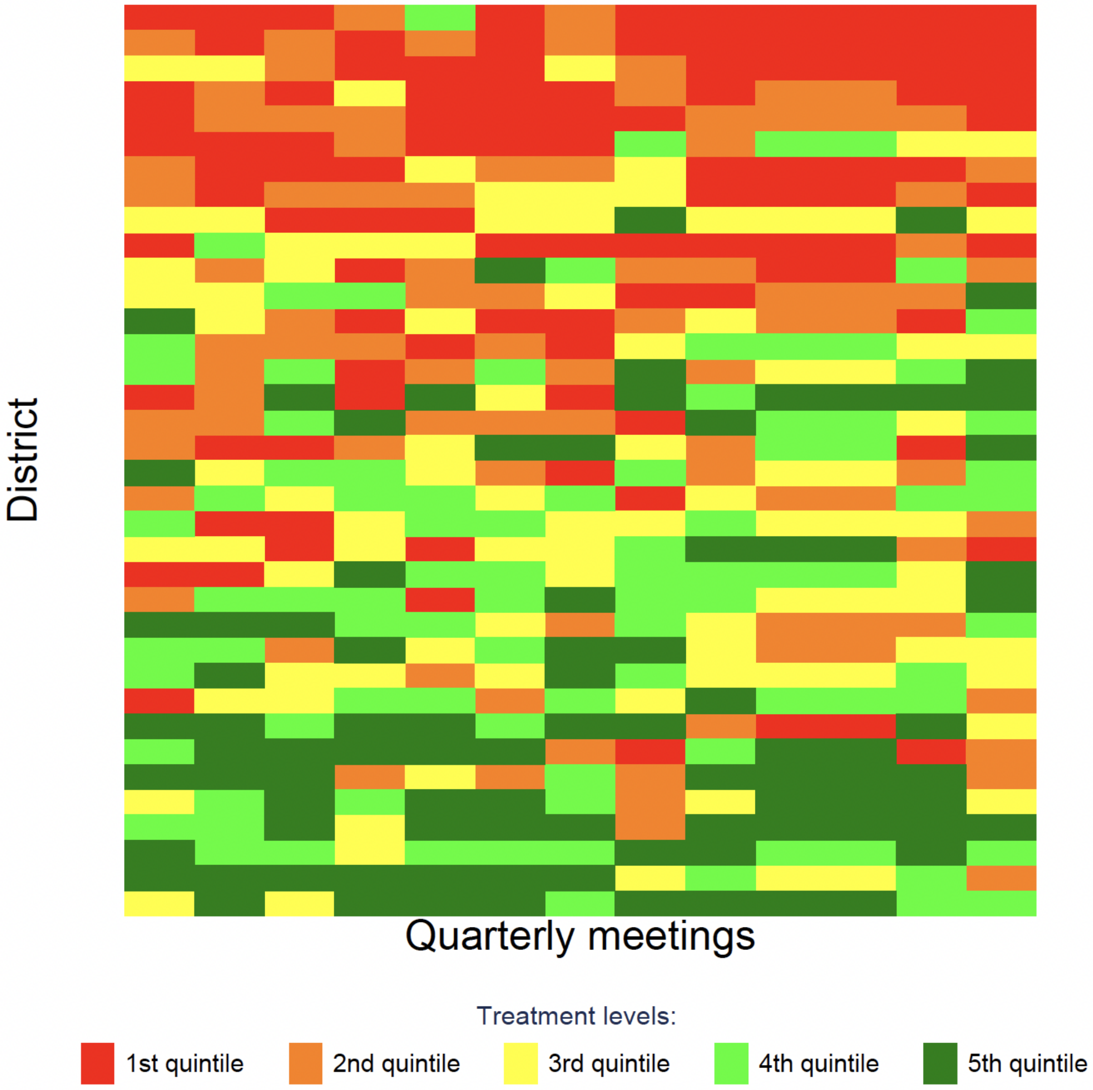
Panel A: School-level variables								
	Mean	**Median**	**Sd**	**N. Obs**	**Mean**	**Median**	**Sd**	**N. Obs**
Number of teachers	4.6	3	3.8	2,305,208	-	-	-	-
Number of students	110	80	102.8	2,307,637	-	-	-	-
Outcomes (%)		No flag				Flag		
Teacher presence	93	100	15	2,092,859	83	100	22	208,003
Student attendance	90	93	12	1,902,131	81	85	17	398,367
Functional facilities	93	100	16	1,876,153	84	100	22	380,310
Math score	87	92	14	824,279	67	67	21	22,274
English score	80	83	17	659,327	65	67	20	187,202
Urdu score	85	89	15	810,174	67	67	20	36,337
Panel B: Markaz-level variables								
	Mean	**Median**	**Sd**	**N. Obs**	**Mean**	**Median**	**Sd**	**N. Obs**
Number of schools	20	15	19	130,357	-	-	-	-
Proportion elementary	80	100	40	130,357	-	-	-	-
Outcomes (%)		No flag				Flag		
Teacher presence	93	94	4.4	95,009	80	83	8	8,653
Student attendance	91	93	6	89,696	80	82	8	13,964
Functional facilities	95	98	11	90,043	81	84	11	13,604
Math score	87	88	6.4	59,974	65	66	5.3	2,119
English score	80	80	6.4	49,363	64	66	5.2	12,730
Urdu score	85	86	6.4	59,295	65	67	5.4	2,798
Panel B: Markaz-level variables								
Outcomes (%)	**Mean**	**Median**	**Sd**	**N. Obs**	**Mean**	**Median**	**Sd**	**N. Obs**
Overall score	94	95	3.8	70	78	78	10	70
New position	7.7	0	2.7	504	8.3	0	2.8	504

Note: Here the unit of observation for outcomes in Panel A is outcome-school-month; in Panel B it is outcome-markaz-month. Outcomes are measured in percentages, with student test scores measured as the percentage of correct answers in standardized tests. A unit is flagged if it receives a flag in the datapack on that outcome in that month. Outcomes in Panel B correspond to the markaz that had elementary schools for which an AEO can be flagged. Panel C reports statistics at the district-quarter level. The "Overall score" variable presents the weighted average of markaz outcomes for a district for the three months before the meeting for those districts ranked at the top or bottom in the respective quarterly meeting. The "New position" variable measures the percentage of districts that enter into the top or bottom in each quarterly meeting.

calendar index—to absorb district linear time trends. Finally, εsmdte accounts for the error term clustered at the markaz level (treatment level). In our main specifications, we stack for four pre-periods and seven post-periods and present robustness for different lengths of periods in the Appendix.

To provide some intuition for our approach, Figure 5 presents the mean evolution of our main outcomes in relative time, anchored on periods of flagging. Solid lines represent schools in maraakiz that are flagged, while dotted lines represent the trajectory of schools in maraakiz that were not flagged. We present two sets of transitions: one that presents statistics using the full sample of schools (the blue

Figure 4: Distribution of Quintiles of District Performance



lines) and one that uses only those schools that are “close” to the threshold for flagging (red lines). We also highlight three periods that correspond to the month in which MEAs collect the data and define the flag, the month in which these are reported to oversight committees and punishments occur, and the period after the flagging events where we assess the impact of treatment.

Table 2: Measures of Variation

School-level variation (sd) by quintiles of overall performance							
Outcomes (%)	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**ALL**	**N. Obs**
Teacher presence	9.8	0.9	0.6	1.5	1.5	7.4	51,534
Student attendance	12.9	1.3	0.7	1.5	1.5	9.5	51,507
Functional facilities	17.3	4.4	1.7	0.6	0.6	16.1	50,501
Math score	5.3	1.1	0.8	1.8	1.8	6.3	37,537
English score	5.8	1.4	1.2	3	3	8.1	37,536
Urdu score	5.5	1.2	0.9	2	2	6.9	37,536

Note: Here the unit of observation for outcomes is presented at the school level. Each quintile is calculated separately based on the mean level of performance for each variable. The table shows the standard deviation of school scores for each variable quintile.

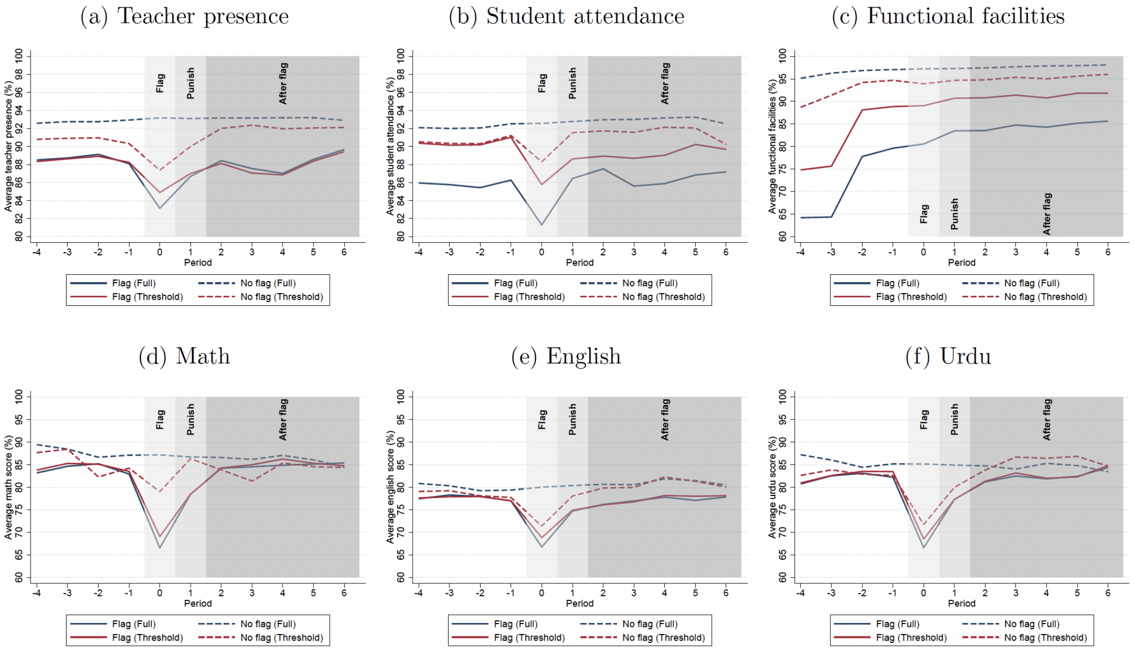
We observe that just before the flagging, treated and control units appear to follow similar paths. In the month of flagging, the average school in a markaz that gets flagged suffers from a shock that contributes to the markaz being selected for treatment.6 Consequently, the treated units would not have followed the same transition as control units in the absence of the treatment. Hence, the parallel trends assumption required for concluding causality would be violated. To address the violation of parallel trends, we follow Rambachan and Roth (2022) and redefine the base period as the one just before the negative transitory shock occurs (relative time - 1).

As can be observed, the transition of outcomes typically reverts to the pre-shock levels. Our preferred specification is to use a regression discontinuity design around the markaz- level flagging thresholds for each outcome to identify samples of schools within an optimal bandwidth either side of the flagging threshold in time 0 (Calonico, Cattaneo, & Farrell, 2020). Because each stack consists of a separate sub-data set, we obtain optimal bandwidths separately for each stack. Hence, the threshold sample is the stacked-threshold sample for each event time. This leads us to study the impact of treatment on two schools that have similar dips in performance, but one that coincides with other schools in its markaz such that it just pushes the markaz into treatment, and one that does not.

Given this approach, γ1 absorbs the effect of the negative transitory shock, while γ2 captures the immediate recovery in the punishment period, just after the shock. AfterFlagte equals 1 for the periods after the recovery, so β would estimate the effect of flagging on school performance after the shock. If the flagging leads to a higher improvement of outcomes on flagged units relative to non-flagged units, β should be positive. That is the core test of the specification. To illustrate the external validity of the results using the sample of schools around the flagging threshold, we also present results for the full set of schools throughout.

[6] This situation is related to an Ashenfelter dip (O. Ashenfelter, 1978; O. C. Ashenfelter & Card, 1984; Heckman & Smith, 1999), which consists of self-selection into the treatment because of a negative shock.

Figure 5: Evolution of school outcomes in relative time - markaz flagging



(a) Teacher presence (b) Student attendance (c) Functional facilities

(d) Math (e) English (f) Urdu

Note: This figure presents the average evolution of schools in flagged (continuous line) and non-flagged (dashed line) maraakiz in the stacked date. Flagging is based on the outcome variable in focus in a particular panel. Blue lines represent the evolution for the full sample, while red accounts for the evolution in a threshold sample that is “close” to the flagging threshold. Relative time is broken into three periods: Flag : the period where information is collected and maraakiz are flagged; Punish: the period where the reports are distributed and oversight meetings are held; After flag : periods after the meeting.

4.2.2 District ranking.

One concern with the above specification is that markaz flagging is less salient when the rest of the district is performing well. We therefore complement our core analysis on markaz flagging with analysis at the district level, as well as with interactions between the two approaches.

Above we noted that in quarterly oversight meetings, districts were ranked according to an overall index of performance of their schools in the prior quarter. Though we are far less empowered to investigate the impact of this ranking, we apply a version of our main specification to being “flagged” as a top- or bottom-performing district on the subsequent performance of schools in that district.

In our analysis of district “flagging,” we stack for four pre-periods and three post-periods, using as event times each month in which a meeting happened. We limit ourselves to three post-periods, considering that district meetings happen each quarter. Flagged units are defined as the schools in districts that were in the bottom or top of the ranking during the meeting in period 0.

In this instance, district rankings do not systematically receive a negative shocks before the meeting and thus do not require corrections for related self-selection and reversion to the mean as in equation 1. However, for consistency we define -1 as the base period and build a threshold sample that consists of the five districts closest to the treated five in the top/bottom to represent a threshold comparison.

We therefore estimate the effect of district ranking on educational outcomes with the following equation:

$$Y_{smdte} = \gamma(Position_{de} \times Meeting_{te}) + \\ \beta(Position_{de} \times AfterMeeting_{te}) + \alpha_{de} + \lambda_{te} + \epsilon_{smdte}$$

Where Positionde equals 1 for schools in bottom/top districts d. Meetingte equals 1 for the period when the quarterly meeting happens (relative time 0), so γ absorbs any immediate effect of the meeting. AfterMeetingte equals 1 for the months after the meeting, so β would estimate the persistent effects of the flagging and is the treatment effect of interest. αde are district fixed effects and λte are time fixed effects. ϵsmdte accounts for the error term clustered at the district level. Interactions between this specification and the above markaz-level specification are natural extensions to these equations.

5. Results

5.1 Markaz flagging

Figure 6 reports the event studies of each variable, flagging the respective outcome, with the y-axis reporting β coefficients in percentage point differences. The blue line represents the full sample, while the red accounts for the threshold sample. The full sample estimations exhibit a dip in the period of flagging, which is almost completely recovered by the period of punishment. The threshold samples track each other more closely by construction.

We see that the negative shock measured in period 0 is almost fully recovered by period 1 in which punishment or treatment occurs. For most of the outcomes, coefficients for the threshold sample are statistically equivalent to zero at the 95 percent level, indicating a zero impact of treatment. For student attendance (panel b), the coefficients on treatment are actually negative, with student attendance taking longer to recover its pre-shock level in treated rather than control schools. However, this is likely an artifact of the fact that treatment schools have a marginally bigger shock in the outcome variable, and they may naturally be assumed to have a longer transition back to equilibrium.

Some exceptions to this overall pattern are in the full sample in the latest periods we examine. These are, of course, when the full sample control schools are most unlike those in treatment, having a longer period without a negative shock. The one exception in the threshold sample is in terms of student attendance after six months, and the improvement is approximately one percentage point. Relative to the numbers in Table 1, this is small in magnitude. Taken together, there is little to no evidence that the high-frequency oversight system has helped improve school or student outcomes.

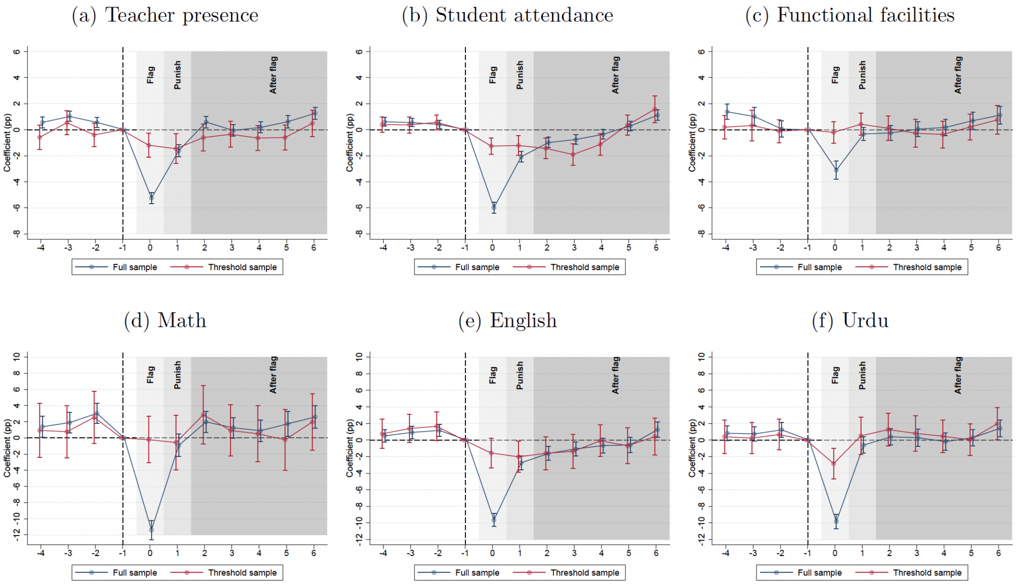
Regarding the identifying assumptions, the event studies show that the pre-trends are not significant or are small in magnitude, particularly for the threshold sample, which suggests the plausibility of parallel trends. As a robustness check concerning the empirical strategy, Figure C1 in the Appendix reports the results of the event studies for a stacked data set with a lower number of post periods to test for the sensibility of the results to define an arbitrary number of periods. Results follow the same trends in both cases, so we can conclude that the results are robust to the parameters of the stacked design. To further test the robustness of the results to the empirical strategy, we rebuild the stacked data set but control for units that have been treated before, so we build a “flagging history fixed effects” model in which we compare units that have had exactly the same treatment history except from the period of analysis in each particular stack. Figure C3 reports the resulting event studies, which follow the same trends as the main specification.

Table 3 presents the quantitative results of estimating equation 1. The first column for each variable reports the full sample results, while the second column reports the threshold sample results. Panel A reports the results for the school outcomes. Except for the threshold sample for functional facilities, there are always negative and significant coefficients in the Flag and Punish periods for flagged units relative to the non-flagged units. The coefficients for both periods represent the first negative shock and the subsequent immediate recovery, which we interpret as reversion to the mean effect. The results highlight the importance of accounting for transitory negative shocks when defining thresholds for performance, as they might be capturing non-permanent affectations to the average behaviour and hence promoting inadequate solutions.

The coefficients for the After flag period (corresponding to β) show that there are no positive and significant effects of treatment. As observed graphically, student attendance and functional facilities show significant but negative results, which might be interpreted as a persistence of the negative shock. Again, these coefficients are small compared to the mean of the dependent variable. These results imply that the oversight scheme had no impacts on school functioning.

Panel B of Table 3 presents the results for the scores variables, and we note that the sample size is smaller here, given the reduced time frame for which we have these measures. The full sample of the three variables reports a significantly worse negative shock, compared to their counterparts in Panel A. We observe the same pattern of results as in Panel A. English scores report small but significant negative coefficients for both samples, which might be attributed to persistence in the negative shock. Overall, the results imply that the oversight scheme had no impacts on student outcomes.

Figure 6: Event Study: Flagging Effect on Performance



Note: This figure presents the results from estimating event studies based on equation 1 using -1 as the base period. The blue line presents results for the full sample, while the red line presents results for the threshold sample. The results are for flagging on the variable in the title of the panel. Error bars at the 95 percent level are presented for each coefficient.

In the Appendix, we assess a range of further robustness checks. We estimate the main results under a different flagging threshold (corresponding to an orange flag) and under a more aggregated flagging structure (corresponding to the tehsil administrative unit). We present additional event study specifications to account for alternative difference-in-differences estimators. To further test the sensitivity of the stacked structure, we plot the average estimate for each coefficient (Flag, Punish, and After flag) from a stacked data set that includes t additional post periods. We test the robustness of the results with respect to changes in the datapack structure and potential time-specific shocks. Finally, we investigate the possibility of the AEOs or other members of the hierarchy anticipating the flagging and shifting their behavior. In all cases, our results are qualitatively the same.

Table 3: Monitoring Effect on Performance: Markaz Flagging

Panel A: School outcomes						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T × Flag	-5.79***	-1.10***	-6.46***	-1.72***	-3.75***	-0.31
	(0.18)	(0.40)	(0.19)	(0.30)	(0.33)	(0.41)
T × Punish	-2.15***	-1.37***	-2.58***	-1.76***	-0.98***	0.31
	(0.20)	(0.50)	(0.17)	(0.35)	(0.21)	(0.41)
T × After flag	-0.052	-0.34	-0.78***	-1.25***	-0.38*	-0.10
	(0.13)	(0.30)	(0.11)	(0.25)	(0.20)	(0.35)
N. of obs.	1,687,227	282,145	1,526,832	249,020	1,554,491	131,139
Mean of Dep. Var. before	90.7	86.4	87.0	86.3	93.5	92.4
R^2	0.064	0.047	0.23	0.11	0.37	0.23
Panel B: Student scores						
Dependent variable:	Math		English		Urdu	
T × Flag	-13.0***	-1.24	-10.3***	-2.53***	-10.5***	-3.18***
	(0.47)	(1.19)	(0.31)	(0.67)	(0.35)	(0.79)
T × Punish	-2.51***	-1.59	-3.45***	-3.01***	-1.33***	0.18
	(0.55)	(1.36)	(0.34)	(0.83)	(0.43)	(0.95)
T × After flag	0.14	0.20	-1.28***	-1.67***	-0.27	0.65
	(0.38)	(1.09)	(0.24)	(0.51)	(0.28)	(0.54)
N. of obs.	506,157	28,900	306,624	73,238	486,326	43,825
Mean of Dep. Var. before	85.6	74.2	73.9	69.7	82.8	70.2
R^2	0.14	0.17	0.091	0.080	0.14	0.16
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

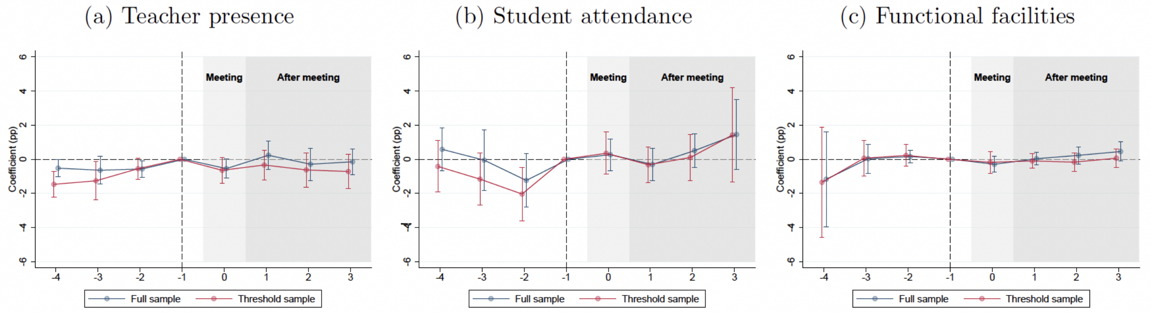
Note: The school is the unit of observation for both panels. T equals 1 for schools in a flagged markaz. Flag equals 1 for the period in which the information is collected and the markaz is flagged, and it is 0 otherwise. Punish is equal to 1 for the period where the reports are distributed and the oversight meeting with corresponding punishment occurs, and it is equal to 0 otherwise. After flag is equal to 1 for periods after the oversight meeting occurs, and it is equal to 0 otherwise. The threshold sample accounts for the schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. Standard errors, clustered by markaz, are in parentheses.* p < 0.10, ** p < 0.05, *** p < 0.01.

5.2 District ranking

We can undertake a similar assessment of the impact of being highlighted as one of the top- or bottom-performing districts overall at quarterly oversight meetings. We restrict our analysis to measures of school functioning given the limited number of quarterly meetings we observe when student test score data is available. Figures 7 and 8 present the results of an event study for both top and bottom performer district-level treatments. The figures illustrate that no pre-periods appear significant, suggesting the plausibility of the parallel trends assumption. There are no significant effects of treatment after a meeting on school functioning.

Table 4 reports the treatment effects. Panel A reports the estimation results for the bottom districts. The results show that only for the threshold sample can we detect a small but significant increase of 1.3 percentage points in student attendance. Overall, however, the results are close to 0. Panel B reports the results for estimation using the schools in top districts as a treatment. The results show that being in a top district leads to a small increase in teacher presence after the quarterly meeting relative to schools in the district just outside of the top five. The coefficients are small in magnitude relative to the mean of the dependent variable before the meeting (91 percent in the full sample and 91.9 percent in the threshold sample). Consequently, there is no evidence supporting significant increases in performance due to centralized monitoring of higher-level managers from the analysis of district-level rankings.

Figure 7: Event Study: Bottom District Effect on Performance

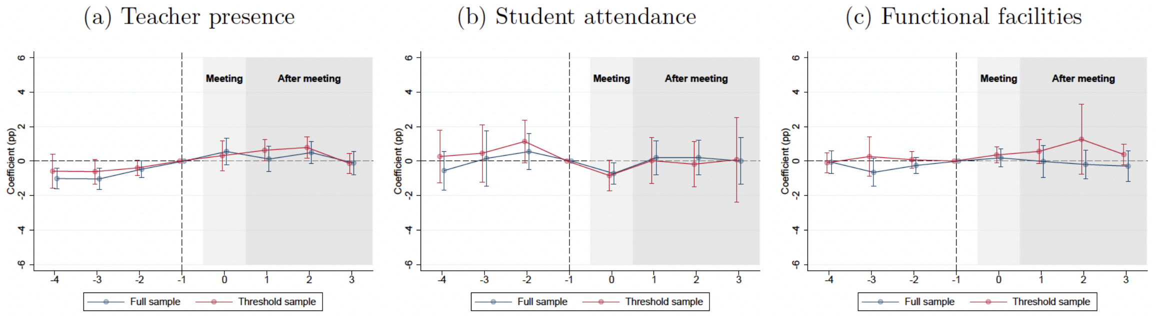

(a) Teacher presence

(b) Student attendance

(c) Functional facilities

Note: This figure presents the results from estimating an event study based on equation 2, using -1 as the base period. The blue line accounts for the results on the full sample, while the red accounts for the results using the threshold sample. Bottom corresponds to the schools in the bottom five districts in the quarterly meeting.

Figure 8: Event Study: Top District Effect on Performance


(a) Teacher presence

(b) Student attendance

(c) Functional facilities

Note: This figure presents the results from estimating an event study based on equation 2, using -1 as the base period. The blue line accounts for the results on the full sample, while the red accounts for the results using the threshold sample. Bottom corresponds to the schools in the bottom five districts in the quarterly meeting.

Table 4: Monitoring Effect on Performance: District Ranking

Panel A: Bottom districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Bottom ×Meeting	-0.11	0.18	0.46	1.28*	-0.045	0.087
	(0.31)	(0.34)	(0.68)	(0.72)	(0.51)	(0.54)
Bottom × After meeting	0.38	0.27	0.72	1.31**	0.48	0.20
	(0.28)	(0.31)	(0.59)	(0.56)	(0.44)	(0.49)
N. of obs.	3,063,894	583,430	3,063,469	583,261	3,009,903	565,933
Mean of Dep. Var. before	91.4	90.1	88.8	86.0	92.5	90.0
R^2	0.025	0.030	0.12	0.15	0.14	0.17
Panel B: Top districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Top × Meeting	1.19***	0.71*	-0.76	-1.32*	0.43	0.29
	(0.30)	(0.38)	(0.51)	(0.73)	(0.30)	(0.28)
Top × After meeting	0.79***	0.82***	0.089	-0.50	0.073	0.66
	(0.25)	(0.23)	(0.31)	(0.68)	(0.42)	(0.46)
N. of obs.	3,111,731	682,496	3,111,137	682,404	3,036,646	672,815
Mean of Dep. Var. before	91.0	91.9	87.4	90.0	91.6	92.7
R^2	0.027	0.026	0.12	0.12	0.14	0.15
Sample	Full	Threshold	Full	Threshold	Full	Threshold
District FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: Bottom is equal to 1 for schools in the bottom five districts on the date of a quarterly meeting. Top equals 1 for the schools in the top five districts on the date of the quarterly meeting. Meeting is equal to 1 for the period in which the quarterly meeting takes place. The threshold sample accounts for the schools in the five districts following the top or bottom. Standard errors, clustered by district, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Such findings are consistent with the descriptive statistics in Panel C of Table 1 and Figure 4, showing that there is little movement into and out of the top quintiles of performance, with corresponding limits on the degree to which they might be motivating.

Despite finding zero overall impacts of flagging at the district level, we tested the hypothesis that the interaction between district-level and markaz-level flagging is positive. We hypothe- sized that the coincidence of flagging at both levels might create greater pressure throughout the hierarchy toward school improvement, leading to a differential increase in performance. We tested this hypothesis by estimating equation 1, including a triple interaction between schools in a bottom or top district in the quarterly meeting and those for which a markaz was also flagged in the month of the meeting. Appendix Table C3 reports the results of the heterogeneity analysis. Panel A reports the results for the bottom districts, while Panel B reports the results for the top districts. Note that the triple interactions for none of the panels, variables, and samples show positive and significant results, suggesting that there is no major interaction effect between flagging district- and markaz-level performance.

5.3 Impacts on the machinery of government

That there were no impacts on school functioning and school outcomes does not mean that the flagging had no impacts on the functioning of government itself. The richness of the administrative data we analyze allows us to investigate bureaucratic activities that we would expect to observe if the bureaucracy were attempting to respond to the flagging.

Physical visits by public officials

First, we explored whether the flagging led to an increase in the effort by the AEOs to improve schools' performance in their respective markaz. From the data collected by the PMIU, it is possible to identify whether the AEO visited each school at the moment of the data collection. The visits are a standard part of the AEOs work program and a mechanism through which to resolve issues that schools face in performing their functions effectively. Such visits are costlier to the AEO in terms of effort, and as such a test of their motivation through the oversight scheme.

Table C5 in the Appendix reports the results of each variable flagging on the measure of effort. The variable of visited schools is measured as a dummy that takes the value of 1 if the school has received a visit from the relevant AEO. Note that the coefficients of the Flag and Punish period account for changes in the probability of receiving a visit, given the negative shock. Still, there are no significant effects for both samples in any variable in such periods. The coefficients for the After flag period are only significant for both samples in the student attendance case, which suggests an increase in the probability between 0.95 percent (full sample) and 2.1 percent (threshold sample) of receiving a visit.

Budget

Beyond physical visits, budgetary resources would seem to be an important margin along which to support struggling schools. We explored the relationship between the extent of flagging and public budget given to schools by aggregating the panel at the year level and calculating for each school the number of times they were in a flagged markaz. We used a panel regression using markaz fixed effects, year fixed effects, and district–time trends to obtain estimates of the effect of the number of times flagged in a year on the amount of funds given for the next year, as well as the expenditures undertaken by the schools in the subsequent year.

Appendix Table C7 shows the results for each flag type on the total amount of funding given by the government in a year to a school and the reported expenses at the school level for the same year. The explanatory variable is the number of times a school was in a flagged markaz in the previous year. Note that for teacher presence, one more flag in the previous year is associated with an increase of 14 percent in the next year expenses and an increase of 2.3 percent in the non-salary funds. However, the rest of the coefficients are broadly small and indistinguishable from zero. A joint test of significance of the full set of coefficients has a p-value lower than 0.01.

Labor market effects

One possible area of de facto management within the power of managers throughout the hierarchy is the ability to transfer, or influence the transfer, of staff below them. As such, a potential mechanism for failure in this setting is the movement of censured staff. Al- ternatively, underperforming schools may be supported more effectively by more effective managers.

We therefore used staffing data to explore whether the oversight scheme had systematic impacts on the movements of public officials throughout the education sector bureaucracy. In particular, we were able to explore the rotation of public officials at the school and district level, measuring rotation as a dummy variable that equals 1 if the public official reported in period t is different from the one reported in t – 1.

We first explored whether the markaz flagging induced higher rotation of head teachers, as AEOs might use the rotation of head teachers as a means of improving school performance. We used equation 1 for each type of flag, using as a dependent variable the rotation of head teachers. We thus estimated the effect of being flagged on the probability of observing head teacher rotation. Appendix Table C6 reports the results. Overall, there are no significant changes in the probability of rotation of head teachers.

For the district level we used equation 2 to observe the rotation in senior managers at bottom-ranked and top-ranked districts. Because the district officer is a district attribute, we aggregated the data at the district level and compared districts in the top or bottom relative to the rest. Panel A of Appendix Table C8 reports the results from being in charge of a bottom-performing district, and Panel B reports the results from being in charge of a top-performing district. We bootstrapped the standard errors because of the low number of observations. No coefficient showed significant results, suggesting that the district flagging system based on rankings does not lead to higher rotation of officers.

Finally, we explored for the district level officers whether being in charge of a top or bottom district was related to a higher- or lower-ranked current position. We obtained data on the current employment of public officers that were in charge of a district at some point between 2011 and 2015. We calculated for each the number of months that they were in charge of a top or a bottom district. We then estimated a simple regression, with bootstrapped standard errors, correlating the ranking of the current employment and the number of months they were in charge.7 Although no coefficient is significant, and we have a relatively small number of observations, we note that the bottom districts are negatively correlated with the rank of the current position of the public official, while the top districts are positively correlated.

Overall, there is no evidence that the oversight scheme induced any substantive impacts in the way government functioned along the key lines of bureaucratic attention, budget, or the public sector labor market. This is, of course, consistent with the null impacts that the scheme had on school outcomes and student test scores.

6. Counterfactual analysis and the persistence of public policy

The results presented above suggest that the high-frequency oversight scheme had no sub- stantive impacts on school functioning or educational outcomes. Why then was the scheme continued for so many years, and interpreted by central managers as such a success?

First, we note that using our approach, the Punjabi Government could have rapidly identified that the oversight component of its reforms was having null to negative impacts on the education service delivery chain. We illustrate this by estimating equation 1 with all data available to the analyst up to month t. In the first month, we apply our empirical approach to data from the first month only, and in the second month, we apply our approach to data from the first two months, and so on. As such, we mimic the analysis that the government itself could have undertaken at any time

[8] We omit the results for the school scores due to the short time series available for these variables

[9] Because we are using only two periods for the estimation, we would be comparing units in maraakiz that might have been more/less flagged in periods before. Then, instead of including markaz fixed effects, we control for flagging history fixed effects.

during the scheme's operation.

Figure C13 in the Appendix plots the after-flag β coefficients for each flagging variable, separating them by full and threshold samples.8 The results are a long string of null or negative coefficients that would have been clearly perceptible within months of the scheme starting.

However, identifying the true effect of the scheme would have required an analytical approach such as that outlined in this paper. Without such an approach, it is feasible that selection for treatment as driven by transitory shocks was not well understood. Rather, recovery from what seems like a transitory shock and regression to the mean may have been interpreted as a causal effect of treatment. To see this, we estimate a reduced version of equation 1 where we compare flagged and non-flagged schools in the Punish period relative to the Flag period.9 Figure C7 in the Appendix shows the evolution of the outcomes between the Flag and Punish periods for the full and threshold sample. We note that in the full sample, the flag trend increases, whereas the non-flagged trend is flat, which might suggest a positive effect of flagging. When accounting for the threshold sample, the trends of both groups appear to be similar.

Appendix Table C4 shows the estimation results. The coefficients suggest that being flagged in the Punish period (relative to the Flag period) is associated with a significant improvement in the performance of all variables relative to the non-flagged units, regardless of the sample. We note then that even if the trend of both groups in the threshold sample appeared to be similar, there still exists a significant positive coefficient for flagged units. Still, the baseline flagging effect (measured by T) shows in all the cases to be significant, negative, and greater than the effect in the Punish period. Consequently, the results demonstrate that observing only the immediate performance after a flag might incorrectly lead the observer to conclude that monitoring has positively impacted performance. Furthermore, the wrong conclusion might persist for the periods after the flag has happened. Figure C8 reports the coefficients from an extended regression to account for what public officials would have observed in the after-flag periods. We note that all the coefficients in the after-flag period are significant and might suggest that the positive effect of flagging persists over time.

7. Using big data to assess the qualities of frontline managers

The results presented above indicate that the high-frequency oversight scheme implemented in Punjab did not have substantive impacts on educational outcomes, but failing to ap- propriately construct a counterfactual analysis might lead the observer to draw incorrect conclusions from the data. Overall, the localized shocks we seem to document are too fast- moving for centralized management. Rather, they are better suited to, and likely subject to, localized management by head teachers.

At the same time, the picture the data paints of the education sector in Punjab is one of systemic bottlenecks to improvements in service delivery. The question this section aims to explore is whether the province's large-scale data effort could have been repurposed toward improved localized management of Punjab's schools' systemic challenges. Lemos et al. (2021) presents evidence that there is significant room for improving the management of schools and substantial variation in how well schools are managed. To what extent then can the data collected by the PMIU identify effective managers (in this case head teachers) and allocate them to where they can be most impactful?

We follow Fenizia (2022) to assess the qualities of head teachers and their impacts on school outcomes and then model the impact of reallocation of head teachers based on these empirical insights as a benchmark to the null effects of the oversight scheme in the use of education management information systems. Such an approach draws on the comparative advantage of large-scale empirical analysis in estimating more permanent parameters of the public sector production function rather than to respond to fast-moving and potentially stochastic shocks.

7.1 Empirical strategy

We follow the procedure proposed by Fenizia (2022) to estimate the effect of a quality-driven change in managers on our outcomes. We start by estimating the following equation:

$$ln(Y_{st}) = \alpha_s + \tau_t + \theta_{m(s,t)} + u_{st}$$

Where In (Y_{st}) is the logarithm of the outcome at the school s for time t. Here αs, τ_t represent school and time fixed effects, while $\theta_{m(s,t)}$ represents head teacher fixed effects. We note that each term captures the variation in the outcome explained by each component. From the equation, we can obtain a predicted value θˆst, representing a measure of how important each head teacher is to explain variances in the outcome, which we interpret as a measure of quality. Given that θˆst is a constant term for each head teacher, we can obtain a shock represented by $\triangle M_s$= θ ˆst −θˆst−1 that represents the change in head teacher quality whenever there is a change of head teacher in a school s so that we can estimate an event study in the form:

$$ln(Y_{st}) = \alpha_s + \sum_{k\neq -1}^{t} (\beta \triangle M_s) + X_{st} + \epsilon_{st}$$

Where β captures the effect of a change in head teacher quality for each period relative to −1 (just before a change of head teacher) and Xst for control variables. Nevertheless, because θˆst is obtained through a two-way fixed effects regression on the outcome ln(Yst) there might be shocks affecting both that might bias the quality measure. Fenizia (2022) exploits the fact that the quality measure averages over all the periods, so the correlation between θst and Yst can be purged for each pair of periods leaving out the data that creates the correlations. Specifically, the process consists in obtaining for each relative time period Δln(Yst) = ln(Yst) – ln(Yst= −1), and $\triangle M_s$, which is obtained through estimating equation 3, without including time 0, −1 (base period) and t (period to be compared with the base period). Then, for each period t/= −1, we would have a regression in the form.

$$\triangle ln(Y_{st}) = \beta_t \triangle \hat{M}_s + X_s + \triangle \epsilon_i$$

The modified event study of equation 5 ensures that each βt is not contaminated by the cor- relation between the quality measure and the outcome. Nevertheless, it requires estimating a separate regression for each event time. In the following section, we present the results for each outcome quality measure so that there is an event study of the impact of head teacher quality for each outcome (using the respective outcome-based quality measure).

7.2 Head teacher quality impact on school performance

Table 5 reports the descriptive characteristics of the data used. Panel A briefly reports statistics at the school level, which, due to requirements of overlap between schools and teachers, is a subset of the total data available. Over the full period we study, schools had an average of 4.6 head teachers, with a median of 5 head teachers and a maximum of 17 head teachers. Among the 52,315 schools in the data, 83 percent had a change in head teacher, which allows us to observe the effect of head teacher changes in most of the education sector of Punjab.

Panel B reports characteristics at the head teacher level. We can observe 219,663 head teachers, most of whom were in charge of only one school. While the maximum number of school switches for a head teacher was 24, only 6.9 percent of head teachers switched schools during our study period. Given the requirements of the method we apply to estimate manager quality, we are restricted to using this subsample of head teachers in what follows.

Table 5: Descriptive Statistics

Panel A: School descriptive variables						
	Mean	**Median**	**Std. Dev**	**Min**	**Max**	**Obs**
Num of head teachers per school (ever)	46	5	2.6	1	17	52.315
% schools with head teacher move (ever)	.83	1	.37	0	1	52,315
Panel B: Head teacher descriptive variables						
	Mean	**Median**	**Std. Dev**	**Min**	**Max**	**Obs**
Num schools per head teacher/date	1	1	.26	1	10	219,663
Num schools per head teacher (ever)	1.1	1	0.4	1	24	219,663
% head teachers move to other school (ever)	.069	0	.25	0	1	219,663

Note: This table reports the characteristics of the unique values of schools and head teachers. The percentage variables are coded as dummy, which takes the value of one when the conditions are met, then the mean captures the percentage of the observations that fulfills the condition. The notation (ever) refers to the complete period, while teacher/date captures the statistics for each particular date.

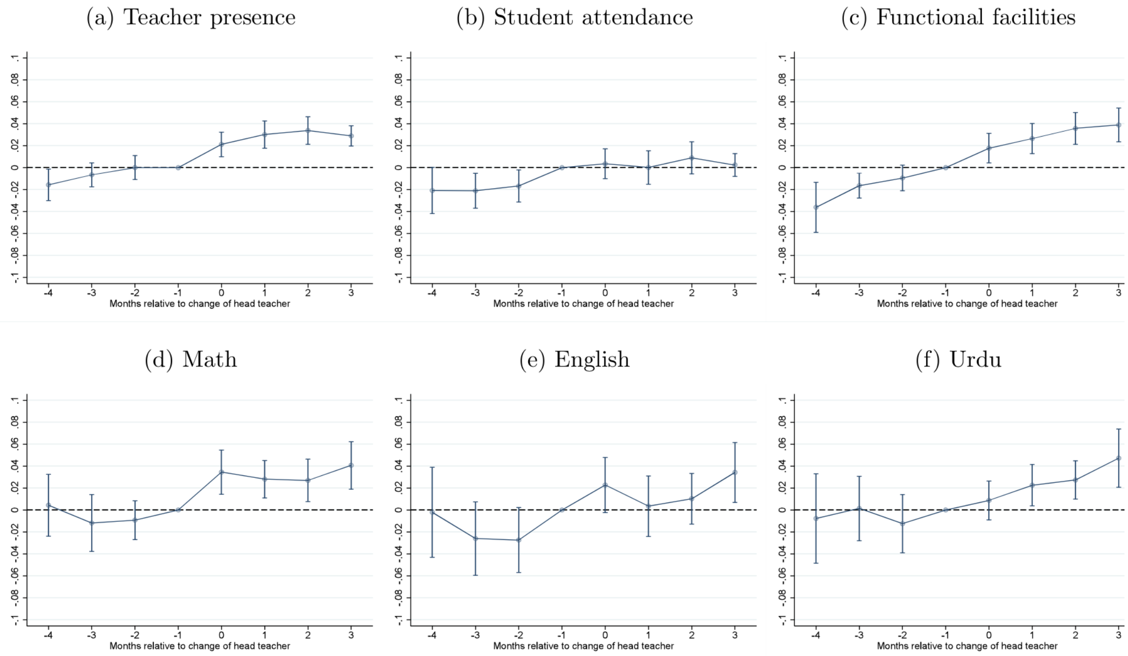
We report the results of the event studies from equation 5 in Figure 9. We control for workload through an interaction between the quality measure and the number of schools for which a head teacher was responsible. Additionally, we control for time-fixed effects and district time trends. The coefficients before the change in head teacher suggest that there are parallel trends so that the outcomes would have evolved similarly in the absence of a change.[10]

The coefficients after the change suggest that for teacher presence, functional facilities, math, and Urdu scores, a higher quality of head teachers represents an immediate improvement. Overall, after a head teacher change, a one standard deviation increase in quality accounts for approximately 3% improvement in the outcomes. The results are much larger for those aspects of school functioning over

[10] We follow Fenizia (2022) and test the validity of the empirical strategy by exploring pre-trends. We divide the head teacher quality by tertiles and plot the coefficient for each tertile before and after the head teacher change. The results are robust as long as the coefficients before the change in head teacher are not significant and the dynamic of the coefficients after the change are symmetric between the last and first tertile.

which the head teacher has direct control-teacher presence and functional facilities-compared with that which s/he does not, such as student attendance. The results suggest that improving the quality of front-line managers might be a mechanism for school improvement.

Figure 9: Event Study: Head Teacher Quality Effect on School Performance



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

(e) English

(f) Urdu

Note: This figure presents the results from estimating an event study based on the method proposed by (Fenizia, 2022). The outcome is in log scale.

Interestingly, according to the predicted values across outcomes for the same head teachers, the managers are not universally talented in all areas of school management. Figure C12 in the Appendix shows the correlation matrix between all of the outcome-quality measures. These findings imply that individual head teachers are most valuable in addressing specific issues, though there appear to be positive correlations between the student scores variables, as might be predicted. This finding showcases one of the powers of large-scale data analysis in the public sector: such analysis can identify the particular talents of specific managers. In turn, these managers can be allocated to schools that match the need for their specific talents, or they can be matched with other managers who may need mentoring in a specific area of school management.

7.3 Modeling effective centralized labor market management

What does the evidence presented in this paper imply for centralized, data-informed public sector management? Our results would seem to imply that the power of empirical analysis is better suited to estimating the structural parameters of the public sector production function and using those for the basis of centralized management.

Section 5 implied that centralized management of routine frontline activities had limited impacts on education outcomes. Section 6 implied that an effective analytics team may have been able to identify these limited effects. Section 7 implies that there is a potential alternative use for the data that the PMIU systematically collects. The centralized data collection enables the identification of talented frontline managers—and the areas in which they are talented—in a way that no decentralized analysis can do. As such, a centralized analytical team focusing on quantifying and managing frontline talent, who then face the stochastic demands of a particular public policy problem—in this case, the management of a specific school—can substantially impact frontline delivery. In this context, we assess how an optimal talent allocation scheme would have impacted the overall productivity of Punjab.

We obtained counterfactual estimates from alternative allocation policies. Following Fenizia (2022) we assume that the productivity of schools depends on the quality of head teachers and schools, both of which the government might observe and which were obtained through the estimation of equation 3. We also assume that the government is capable of effectively allocating head teachers to schools. The counterfactual estimates are then obtained by matching different head teachers with schools and obtaining the new hypothetical value of the variable under the new allocation, and then comparing it with the original.[11]

Table 6 reports the percentage difference between the hypothetical allocation and the origi- nal. We focus on reallocating head teachers in schools performing below the median for each variable and test for two policies: i) optimal allocation, consisting in matching schools with teachers based on ranking (good schools with good head teachers); and ii) firing the bottom 20 percent of head teachers and replacing them with median-quality head teachers. Addi- tionally, we report the productivity gains in the overall sample of schools (including those for which there was no reallocation) and only in the schools below the median of performance (those with reallocation). We note that regardless of the counterfactual policy, there are great gains in productivity for the schools performing below the median when allocating better head teachers in the specific areas in which schools were underperforming.

[11] Considering a transformation of equation 3 as $Y\hat{}s = \exp(\alpha\hat{}s + \theta\ \hat{}\ )$, we can obtain new values $Y\hat{}s$ for each new distribution of the estimated parameters.

Table 6: Productivity Gains from Hypothetical Head Teachers Allocation Policies

Outcome	Optimal allocation		Replacing bottom 20% with the median	
Teacher presence	4.3%	19.3%	1.6%	7.1%
Student attendance	4.1%	8.4%	1.8%	3.7%
Functional facilities	2.4%	71.5%	0.7%	20.8%
Math score	6.4%	13.6%	3.1%	6.5%
English score	8.3%	21.5%	3.2%	8.3%
Urdu score	6.3%	18.7%	2.8%	8.2
Sample of schools	All	Below the median	All	Below the median

Note: This table reports the percentage difference between the hypothetical policy allocation against the original productivity. Optimal allocation consists in matching (by order) good head teachers with good schools. The second policy consist in firing the bottom 20 percent of head teachers in terms of quality and replacing them with the median head teacher (assuming the government can effectively replace them). The results are reported for the sample of all schools, and for the sample of schools performing below the median of each variable.

8. Discussion

This paper undertakes analysis of administrative data from Punjab province in Pakistan to assess the efficacy of centralized monitoring of school outcomes on the quality of ser- vices delivered. In particular, we focus on a flagging system based on minimum service thresholds. We find that the system had no appreciable impact on the trajectory of school outcomes across any area of its focus. This is in contrast to the significant impact that we estimate frontline managers (head teachers) can have on school outcomes across the same areas of focus. Thus, the evidence from analysis of Punjab's own administrative data is that centralized management approaches struggle to effectively manage unpredictable and varying delivery environments but they may be able to use large-data to estimate important structural elements of the public sector production function.

The pattern observed in the Punjab data is instructive. It clearly records a negative tran- sitory shock in an outcome—say, teacher presence—that leads to the markaz being flagged. Immediately after the shock, there is an improvement in the measured outcome that corre- sponds with a reversion to the mean effect. Once the flag is published, that reversion to the mean continues, and thus, a simple analysis might perceive the flag as having had an effect. However, when a proper counterfactual is compared with the flagged jurisdiction, flagging shows no effects on any of the educational variables we study. Both environments revert to the mean in the same way. Therefore, we argue that the flagging does not generate more effort from targeted public officials toward improvement than those in a carefully constructed counterfactual.

Furthermore, we find no quantitatively significant impacts of the flagging at any point in the hierarchy. We explore whether those districts (a higher level of aggregation than the core analysis) that were flagged as poorly performing in the meetings with senior officials have a differential trajectory in their aggregate educational outcomes compared to those districts that were not flagged. We see no impact on the relative ranking of districts over the study period. The worst-performing districts remain ranked at the bottom of the reports, and the best stay ranked as the best. Thus, the relative inequality in educational opportunities across the province was unchanged by the oversight approach.

The results suggest that the high-frequency monitoring system in Punjab failed to generate improvements in the education sector of Punjab. There are three key corollaries to this finding. First, despite having a granular and widespread measurement system in place, the absence of a counterfactual analysis led to the continuation and expansion of the program over a multi-year period. Using data only from the early periods, we find that it was quickly apparent that it was having no effects. Thus, the results presented in this paper were available at an early stage of the program to senior managers, had the right analytical team been in place to assist them with interpretation. Second, there do not seem to have been any costs to the public officials involved in poorly performing schools or districts in terms of their career progression. This may indicate that the medium-term incentives for performance in the education sector of Punjab are not effectively tied to impacts on school outcomes. Third, there may have been a wider learning benefit to the system, but it was not one that differentially improved the performance of the weakest performers as defined by the rankings at the start of the study period. Rather, the relative rankings of school quality stayed very stable. A threshold-based approach to performance measurement also does not seem to be the most relevant for a system that aims to maximize learning. Alternative reporting based on the same data may have captured relative progress better.

Our paper provides a detailed evaluation of the concerns with accountability systems that have been debated in the literature (Kane & Staiger, 2002; Besley & Coate, 2003; Bardhan, 2002; B´o, Finan, Li, & Schechter, 2021). Centralized management may not have the capacity to drive management changes throughout a public sector hierarchy required by an oversight approach to government. However, re-purposing the data that underlies an oversight ap- proach to government for analytical purposes related to structural determinants of public sector effectiveness has much greater promise (Lang, 2010; Staiger & Rockoff, 2010).

References

Ali, A. J., Fuenzalida, J., G´omez, M., & Williams, M. J. (2021). Four lenses on people management in the public sector: An evidence review and synthesis. Oxford Review of Economic Policy, 37 (2), 335–366. Retrieved from https://ideas.repec.org/a/ oup/oxford/v37y2021i2p335-366..html

Ash, E., & MacLeod, W. B. (2015). Intrinsic motivation in public service: Theory and evi- dence from state supreme courts. The Journal of Law and Economics, 58 (4), 863–913. Retrieved from https://doi.org/10.1086/684293 doi: 10.1086/684293
Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. The Review of Economics and Statistics, 47–57.

Ashenfelter, O. C., & Card, D. (1984). Using the longitudinal structure of earnings to estimate the effect of training programs. National Bureau of Economic Research.

Ashraf, N., Bandiera, O., & Jack, B. K. (2014). No margin, no mission? a field experiment on incentives for public service delivery. Journal of Public Economics, 120 , 1–17.

Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates? Journal of Financial Economics, 144 (2), 370–395. Bandiera, O., Best, M. C., Khan, A. Q., & Prat, A. (2021, 08). The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats*. The Quarterly Journal
of Economics, 136 (4), 2195–2242. Retrieved from https://doi.org/10.1093/qje/ qjab029 doi: 10.1093/qje/qjab029

Banerjee, A., Chattopadhyay, R., Duflo, E., Keniston, D., & Singh, N. (2021, February). Improving police performance in rajasthan, india: Experimental evidence on incentives, managerial autonomy, and training. American Economic Journal: Economic Policy, 13 (1), 36–66. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pol
.20190664 doi: 10.1257/pol.20190664

Banerjee, A. V., Duflo, E., & Glennerster, R. (2008). Putting a band-aid on a corpse: Incentives for nurses in the indian public health care system. Journal of the Euro- pean Economic Association, 6 (2-3), 487-500. Retrieved from https://onlinelibrary
.wiley.com/doi/abs/10.1162/JEEA.2008.6.2-3.487 doi: https://doi.org/10.1162/ JEEA.2008.6.2-3.487

Barber, M. (2013). The good news from pakistan (Tech. Rep.). Re- form, London. https://assets.website-files.com/59ca37d5fcfbf3000197aab3/ 5be1df67f395d780786441d8 Pakistan%20final.pdf.

Barber, M. (2021). Accomplishment: How to achieve ambitious and challenging things. London: Penguin UK.

Bardhan, P. (2002, December). Decentralization of governance and development. Journal of Economic Perspectives, 16 (4), 185-205. Retrieved from https://www.aeaweb.org/ articles?id=10.1257/089533002320951037 doi: 10.1257/089533002320951037

Bau, N., & Das, J. (2020, February). Teacher value added in a low-income country. Amer- ican Economic Journal: Economic Policy, 12 (1), 62-96. Retrieved from https:// www.aeaweb.org/articles?id=10.1257/pol.20170243 doi: 10.1257/pol.20170243

B´ekir, I., Harbi, S. E., Grolleau, G., Mzoughi, N., & Sutan, A. (2016). The impact of monitoring and sanctions on cheating: experimental evidence from tunisia. Managerial and Decision Economics, 37 (7), 461–473.

Belot, M., & Schr¨oder, M. (2016). The spillover effects of monitoring: A field experiment. Management Science, 62 (1), 37–45.

Bertrand, M., Burgess, R., Chawla, A., & Xu, G. (2020). The glittering prizes: Career incentives and bureaucrat performance. The Review of Economic Studies, 87 (2), 626– 655.

Besley, T., Burgess, R., Khan, A., & Xu, G. (2022). Bureaucracy and development. Annual Review of Economics, 14 (1), 397-424. Retrieved from https://doi.org/10.1146/annurev-economics-080521-011950 doi:10.1146/annurev-economics-080521-011950

Besley, T., & Coate, S. (2003, December). Centralized versus decentralized provision of local public goods: a political economy approach. Journal of Public Economics, 87 (12), 2611–2637. Retrieved from https://doi.org/10.1016/s0047-2727(02)00141-x doi: 10.1016/s0047-2727(02)00141-x

Bo, E. D., Finan, F., Li, N. Y., & Schechter, L. (2021). Information technology and government decentralization: Experimental evidence from paraguay. Econometrica, 89 (2), 677–701. Retrieved from https://doi.org/10.3982/ecta17497 doi: 10.3982/ ecta17497

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2012). Estimating the effect of leaders on public sector productivity: The case of school principals (Tech. Rep.). National Bureau of Economic Research.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time peri- ods. Journal of Econometrics, 225 (2), 200–230.

Callen, M., Gulzar, S., Hasanain, A., Khan, M. Y., & Rezaee, A. (2020). Data and policy decisions: Experimental evidence from pakistan. Journal of Development Economics, 146 , 102523.

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. The Econometrics Journal , 23 (2), 192–210.

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. The Quarterly Journal of Economics, 134 (3), 1405–1454.

Center for Global Development. (2022). Schooling for all: Feasible strategies to achieve universal education (Tech. Rep.). Center for Global Development.

Chaudhry, R., & Tajwar, A. W. (2021). The punjab schools reform roadmap: A medium- term evaluation. In F. M. Reimers (Ed.), Implementing deeper learning and 21st cen- tury education reforms: Building an education renaissance after a global pandemic (pp. 109–128). Cham: Springer International Publishing. Retrieved from https:// doi.org/10.1007/978-3-030-57039-2 5 doi: 10.1007/978-3-030-57039-2 5
\
Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014, September). Measuring the im- pacts of teachers ii: Teacher value-added and student outcomes in adulthood. Amer- ican Economic Review , 104 (9), 2633-79. Retrieved from https://www.aeaweb.org/ articles?id=10.1257/aer.104.9.2633 doi: 10.1257/aer.104.9.2633

Coelli, M., & Green, D. A. (2012). Leadership effects: school principals and stu- dent outcomes. Economics of Education Review , 31 (1), 92-109. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775711001488 doi: https://doi.org/10.1016/j.econedurev.2011.09.001

Crawfurd, L., & Rolleston, C. (2020). Long-run effects of teachers in developing coun- tries. Review of Development Economics, 24 (4), 1279-1299. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/rode.12717 doi: https://doi.org/ 10.1111/rode.12717

Dal Bo, E., Finan, F., & Rossi, M. A. (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. The Quarterly Journal of Economics, 128 (3), 1169–1218.

De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review , 110 (9), 2964–96.

Deserranno, E. (2019, January). Financial incentives as signals: Experimental evidence from the recruitment of village promoters in uganda. American Economic Journal: Applied Economics, 11 (1), 277-317. Retrieved from https://www.aeaweb.org/articles?id= 10.1257/app.20170670 doi: 10.1257/app.20170670

Deserranno, E., Leon, G., & Kastrau, P. (2022). Promotions and productivity: The role of meritocracy and pay progression in the public sector (Tech. Rep.). Working Paper.

Dhaliwal, I., & Hanna, R. (2017). The devil is in the details: The successes and limitations of bureaucratic reform in india. Journal of Development Economics, 124 , 1–21.

Dickinson, D., & Villeval, M.-C. (2008). Does monitoring decrease work effort?: The comple- mentarity between agency and crowding-out theories. Games and Economic behavior , 63 (1), 56–76.

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. The Journal of Human Resources, 37 (4), 696–727. Retrieved 2023-02-04, from http://www.jstor.org/stable/3069614

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. American Economic Review , 102 (4), 1241–78.
Falk, A., & Kosfeld, M. (2006). The hidden costs of control. American Economic Review, 96 (5), 1611–1630.

Fenizia, A. (2022). Managers and productivity in the public sector. Econometrica, 90 (3), 1063–1084.

Finan, F., Olken, B. A., & Pande, R. (2015). The personnel economics of the state. Handbook of Economic Field Experiments.

Glewwe, P., & Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. Handbook of the Economics of Education, 2 , 945–1017.
Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing.

Journal of Econometrics, 225 (2), 254–277.

Hanushek, E. A., & Rivkin, S. G. (2006). Chapter 18 teacher quality. In E.

Hanushek & F. Welch (Eds.), (Vol. 2, p. 1051-1078). Elsevier. Retrieved from https://www .sciencedirect.com/science/article/pii/S1574069206020186 doi: https://doi.org/10.1016/S1574-0692(06)02018-6

Heckman, J. J., & Smith, J. A. (1999). The pre-programme earnings dip and the deter- minants of participation in a social programme. implications for simple programme evaluation strategies. The Economic Journal , 109 (457), 313–348.
Honig, D. (2021). Supportive management practice and intrinsic motivation go to- gether in the public service. Proceedings of the National Academy of Sciences,

118 (13), e2015124118. Retrieved from https://www.pnas.org/doi/abs/10.1073/ pnas.2015124118 doi: 10.1073/pnas.2015124118

Hussain, I. (2015, February). Subjective performance evaluation in the public sector evidence from school inspections. Journal of Human Resources, 50 (1), 189–221.

Kane, T. J., & Staiger, D. O. (2002, November). The promise and pitfalls of us- ing imprecise school accountability measures. Journal of Economic Perspectives, 16 (4), 91–114. Retrieved from https://doi.org/10.1257/089533002320950993 doi: 10.1257/089533002320950993

Khan, A. Q., Khwaja, A. I., & Olken, B. A. (2019, January). Making moves mat- ter: Experimental evidence on incentivizing bureaucrats through performance-based postings. American Economic Review , 109 (1), 237-70. Retrieved from https:// www.aeaweb.org/articles?id=10.1257/aer.20180277 doi: 10.1257/aer.20180277

Lang, K. (2010, September). Measurement matters: Perspectives on education policy from an economist and school board member. Journal of Economic Perspectives, 24 (3), 167-
82. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jep.24.3.167
doi: 10.1257/jep.24.3.167
Leaver, C., Lemos, R., & Scur, D. (2019). Measuring and explaining management in schools

: New approaches using public data (Policy Research Working Paper). World Bank Group. Retrieved from https://openknowledge.worldbank.org/handle/10986/ 32662

Leaver, C., Ozier, O., Serneels, P., & Zeitlin, A. (2021, July). Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from rwandan primary schools. American Economic Review , 111 (7), 2213-46. Re- trieved from https://www.aeaweb.org/articles?id=10.1257/aer.20191972 doi: 10.1257/aer.20191972

Lemos, R., Muralidharan, K., & Scur, D. (2021, January). Personnel management and school productivity: Evidence from india (Working Paper No. 28336). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w28336 doi: 10.3386/w28336

Liebowitz, D. D., & Porter, L. (2019, July). The effect of principal behaviors on student, teacher, and school outcomes: A systematic review and meta-analysis of the empirical literature. Review of Educational Research, 89 (5), 785–827. Retrieved from https:// doi.org/10.3102/003465431986613 doi: 10.3102/0034654319866133

Malik, R., & Bari, F. (2022). Improving service delivery via top-down data-driven account- ability: Reform enactment of the education road map in pakistan (Tech. Rep.). Working Paper.

Mbiti, I. M. (2016, September). The need for accountability in education in developing countries. Journal of Economic Perspectives, 30 (3), 109-32. Retrieved from https:// www.aeaweb.org/articles?id=10.1257/jep.30.3.109 doi: 10.1257/jep.30.3.109

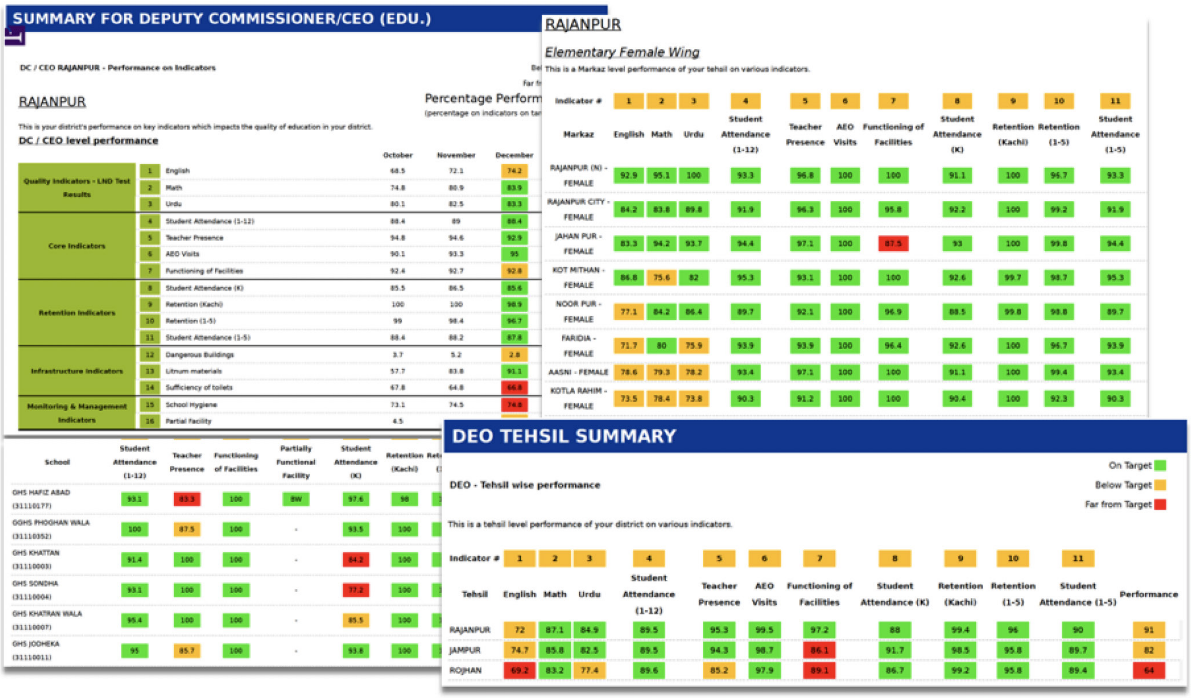
Muralidharan, K., & Niehaus, P. (2017, November). Experimentation at scale. Journal of Economic Perspectives, 31 (4), 103-24. Retrieved from https://www.aeaweb.org/ articles?id=10.1257/jep.31.4.103 doi: 10.1257/jep.31.4.103

Muralidharan, K., Niehaus, P., Sukhtankar, S., & Weaver, J. (2021). Improving last-mile service delivery using phone-based monitoring. American Economic Journal: Applied Economics, 13 (2), 52–82.

Muralidharan, K., & Singh, A. (2020). Improving public sector management at scale? experimental evidence on school governance india (Tech. Rep.). National Bureau of Economic Research.
Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. Journal of Political Economy, 119 (1), 39–77.
Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in indonesia.
Journal of political Economy, 115 (2), 200–249.
Rambachan, A., & Roth, J. (2022). A more credible approach to parallel trends (Tech. Rep.).
Working Paper.
Rasul, I., & Rogger, D. (2018). Management of bureaucrats and public service delivery: Evidence from the nigerian civil service. The Economic Journal , 128 (608), 413-446. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12418 doi: https://doi.org/10.1111/ecoj.12418
Rasul, I., Rogger, D., & Williams, M. J. (2020, 11). Management, Organizational Perfor- mance, and Task Clarity: Evidence from Ghana's Civil Service. Journal of Public Ad- ministration Research and Theory , 31 (2), 259-277. Retrieved from https://doi.org/ 10.1093/jopart/muaa034 doi: 10.1093/jopart/muaa034
Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73 (2), 417–458.
Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. American economic review , 94 (2), 247–252.
School Education Department. (2018). Annual school census (Tech. Rep.). Government of Punjab.
Staiger, D. O., & Rockoff, J. E. (2010, September). Searching for effective teachers with imperfect information. Journal of Economic Perspectives, 24 (3), 97-118. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jep.24.3.97 doi: 10.1257/ jep.24.3.97
Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, 225 (2), 175–199.
Vivalt, E. (2020, 09). How Much Can We Generalize From Impact Evaluations? Journal of the European Economic Association, 18 (6), 3045-3089. Retrieved from https:// doi.org/10.1093/jeea/jvaa019 doi: 10.1093/jeea/jvaa019
World Bank Group. (2018). World development report 2019: Learning to realize education's promise (Tech. Rep.). World Bank Publications.
World Bank Group. (2020). How to improve education outcomes most efficiently? a com- parison of 150 interventions using the new learning-adjusted years of schooling metric.

A. Appendix: Datapack images

Figure 4: Distribution of Quintiles of District Performance



SUMMARY FOR DEPUTY COMMISSIONER/CEO (EDU.)

DC / CEO RAJANPUR - Performance on Indicators

RAJANPUR

Percentage Perform

(percentage on indicators on tar

This is your district's performance on key indicators which impacts the quality of education in your district.

DC / CEO level performance

			October	November	December
Quality Indicators - LND Test Results	1	English	68.5	72.1	74.2
	2	Math	74.8	80.9	83.9
	3	Urdu	80.1	82.5	83.3
Core Indicators	4	Student Attendance (1-12)	88.4	89	88.4
	5	Teacher Presence	94.8	94.6	92.9
	6	AEO Visits	90.1	93.3	95
	7	Functioning of Facilities	92.4	92.7	92.8
Retention Indicators	8	Student Attendance (K)	85.5	86.5	85.6
	9	Retention (Kachi)	100	100	98.9
	10	Retention (1-5)	99	98.4	96.7
	11	Student Attendance (1-5)	88.4	88.2	87.8
Infrastructure Indicators	12	Dangerous Buildings	3.7	5.2	2.8
	13	Litnum materials	57.7	83.8	91.1
	14	Sufficiency of toilets	67.8	64.8	66.8
Monitoring & Management Indicators	15	School Hygiene	73.1	74.5	74.8
	16	Partial Facility	4.5		

School	Student Attendance (1-12)	Teacher Presence	Functioning of Facilities	Partially Functional Facility	Student Attendance (K)	Retention (Kachi)
GHS HAFIZ ABAD (31110177)	93.1	83.3	100	BW	97.6	98
GGHS PHOGHAN WALA (31110352)	100	87.5	100	-	93.5	100
GHS KHATTAN (31110003)	91.4	100	100	-	84.2	100
GHS SONDHA (31110004)	93.1	100	100	-	77.2	100
GHS KHATRAN WALA (31110007)	95.4	100	100	-	85.5	100
GHS JODHEKA (31110011)	95	85.7	100	-	93.8	100

RAJANPUR

Elementary Female Wing

This is a Markaz level performance of your tehsil on various indicators.

Indicator #	1	2	3	4	5	6	7	8	9	10	11
Markaz	English	Math	Urdu	Student Attendance (1-12)	Teacher Presence	AEO Visits	Functioning of Facilities	Student Attendance (K)	Retention (Kachi)	Retention (1-5)	Student Attendance (1-5)
RAJANPUR (N) - FEMALE	92.9	95.1	100	93.3	96.8	100	100	91.1	100	96.7	93.3
RAJANPUR CITY - FEMALE	84.2	83.8	89.8	91.9	96.3	100	95.8	92.2	100	99.2	91.9
JAHAN PUR - FEMALE	83.3	94.2	93.7	94.4	97.1	100	87.5	93	100	99.8	94.4
KOT MITHAN - FEMALE	86.8	75.6	82	95.3	93.1	100	100	92.6	99.7	98.7	95.3
NOOR PUR - FEMALE	77.1	84.2	86.4	89.7	92.1	100	96.9	88.5	99.8	98.8	89.7
FARIDIA - FEMALE	71.7	80	75.9	93.9	93.9	100	96.4	92.6	100	96.7	93.9
AASNI - FEMALE	78.6	79.3	78.2	93.4	97.1	100	100	91.1	100	99.4	93.4
KOTLA RAHIM - FEMALE	73.5	78.4	73.8	90.3	91.2	100	100	90.4	100	92.3	90.3

DEO TEHSIL SUMMARY

DEO - Tehsil wise performance

On Target 🟩
Below Target 🟨
Far from Target 🟥

This is a tehsil level performance of your district on various indicators.

Indicator #	1	2	3	4	5	6	7	8	9	10	11	
Tehsil	English	Math	Urdu	Student Attendance (1-12)	Teacher Presence	AEO Visits	Functioning of Facilities	Student Attendance (K)	Retention (Kachi)	Retention (1-5)	Student Attendance (1-5)	Performance
RAJANPUR	72	87.1	84.9	89.5	95.3	99.5	97.2	88	99.4	96	90	91
JAMPUR	74.7	85.8	82.5	89.5	94.3	98.7	86.1	91.7	98.5	95.8	89.7	82
ROJHAN	69.2	83.2	77.4	89.6	85.2	97.9	89.1	86.7	99.2	95.8	89.4	64

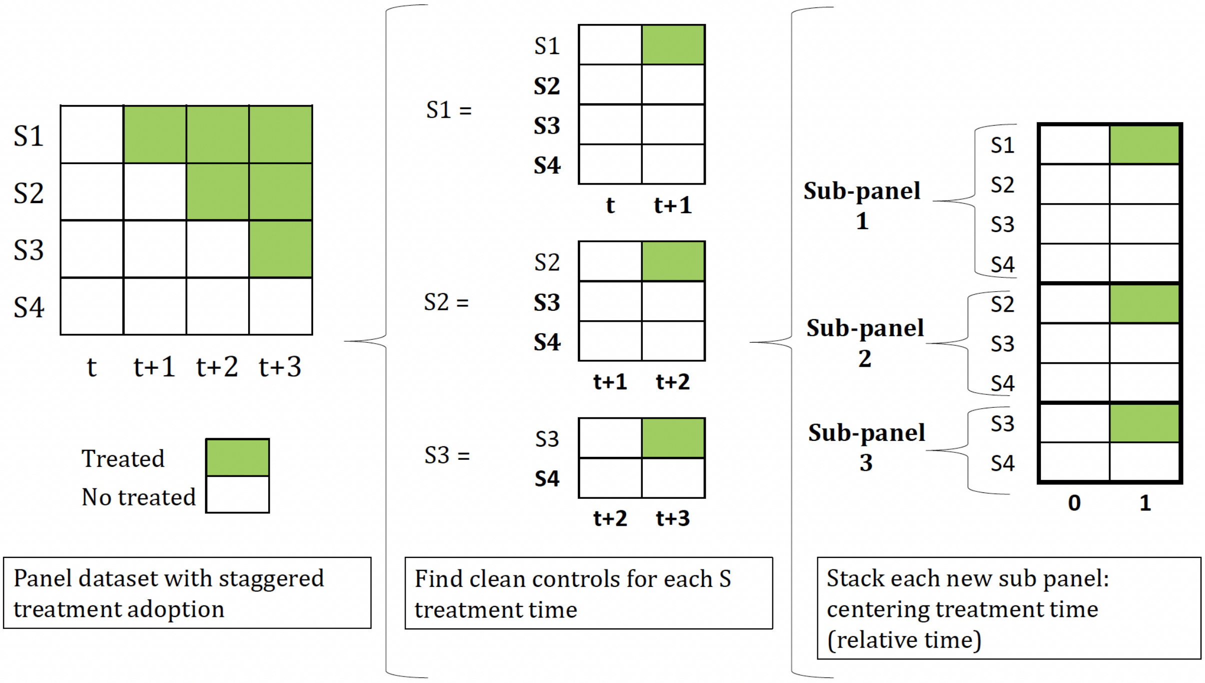
B. Appendix: Data description and stacking process

Stacking process

Figure B1 intends to briefly describe the stacking procedure followed by Cengiz et al. (2019) to center a treatment in a staggered treatment adoption setting, to eliminate the bias pro- duced by the varying timing nature of the treatment. We assume a panel with four subjects, S1, S2, S3, S4, and four periods, t, t + 1, t + 2, t + 3. Each row/column of the matrix cor- responds to a subject/period treatment status. Each time it turns green, the respective subject starts to be treated. Note that except for S4, which is never treated, all of the other subjects adopt the treatment in different periods of time.

The main objective of the stacking is to find for each treated unit a set of clean controls that have not been treated across a period of time. Consequently, we must arbitrarily choose some pre- and post-periods where some sets of control units are not treated for the periods after the treatment of a particular unit. For this example, we take one period before the treatment and the period of treatment adoption to find for each treated unit the set of controls. For S1 we consider the times t and t + 1. Note that until t + 1, no other subject has been treated. Hence, S2, S3, S4 compose the set of clean controls for S1. Note that although S2 and S3 will be treated until t + 1, they have not been exposed. For S2 we consider time t + 1 and t + 2, for which S3, S4 are not treated, then they compose the set of clean controls for S2. In the case of S3, we consider times t + 2 and t + 3, for which the only unit not yet treated is S4, which consists of the set of clean controls for S3.

Figure B1: Stacking Process


S1
S2
S3
S4
t
t+1
t+2
t+3
Treated
No treated
Panel dataset with staggered treatment adoption
S1 =
S2 =
S3 =
Find clean controls for each S treatment time
Sub-panel 1
Sub-panel 2
Sub-panel 3
0
1
Stack each new sub panel: centering treatment time (relative time)

As observed in the figure, for each treated unit, we were able to build a two-period panel data set, for which each panel had its own clean controls. The last step consisted of assigning a unique identifier for each panel so that we could stack them together and keep track of each one separately. Note that although all the panels were built based on different time periods, one can normalize in relative time, such that the treatment is always happening simultaneously. Hence, no bias from treatment timing adoption appears from estimating through two-way fixed effects. Additionally, note that the same unit can appear as a treated or control unit at different times in the panel. Consequently, for estimations purposes, the unit and time fixed effects must interact with panel-unique identifiers to account for repeated units and differences in relative time origins.

Drawbacks may arise from using the stacked data, particularly in our setting. Because the underperformance of maraakiz/districts might change between dates, the treatment can be turned on/off. Consequently, the arbitrary decision of choosing post-periods leads us to assume that the unit remains treated, even if it is turned off (or turned back on) in the sense of the stacking approach we take. In such a case, we might be losing information about the treatment. Considering this limitation, we present results for different time ranges (see Figures 6 and C1). Additionally, although the process described here allows us to account for the time-varying nature of the treatment, we still obtained the results through two-way fixed effects, which in turn might still be biased under heterogeneous responses to the treatment (Goodman-Bacon, 2021). We therefore tested additional estimators to check the robustness of the results.

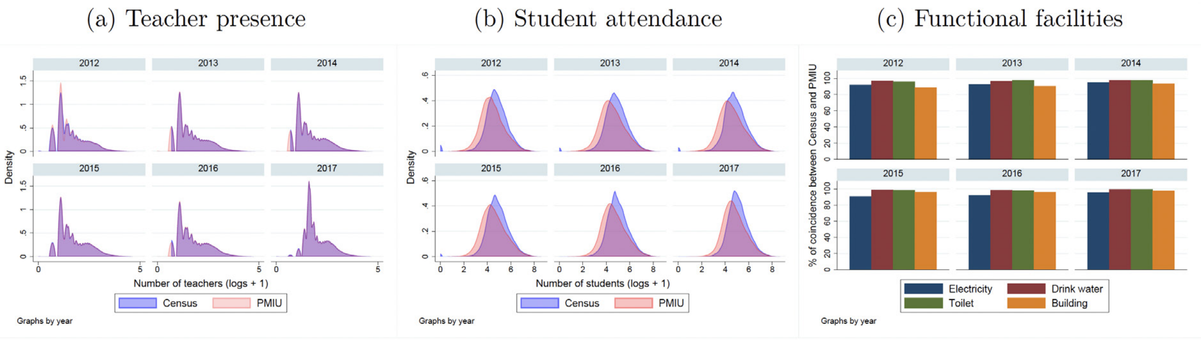
Compliance

The system had high compliance rates. In this regard, we have datapack reports for 60 months, from December 2011 to May 2018, which account for 100 percent of possible re- porting (without including June, July, and August, for which no reports were generated). As long as all possible reports were produced, the education authorities had available the data to flag units and enforce the respective punishments. As a result, we have monthly monitoring data for all the generated datapacks.

To test the level of compliance with the quality of the monthly data, we compared it against the annual census of schools in Punjab, which was independently collected. The annual census was collected in October of every year; we therefore compared it against the October monthly data used for the datapacks. In particular, both data sources reported information about the number of teachers posted, students enrolled, and the functionality of the school infrastructure. Figure B2 compares the distribution of the variables in both sources. Panels (a) and (b) plot the distribution of the log-transformed teacher presence and student attendance from 2012 to 2017. It can be observed that both overlap, suggesting that even if there were differences, most of the population was mapped consistently between the monthly data and the census data. Panel (c) plots the percentage of schools where the report of functionality in different infrastructures coincides; in both sources, the school reported the infrastructure as functional or not functional. The figure suggests that for all years, the percentage of coincidence is near 100 percent.

Alternatively, Figure B3 plots the distribution of the differences in the reporting between PMIU monthly data and census data. Panel (a) plots the distribution of the differences for teacher presence. Despite having some tails on both sides, most of the reports correspond to a difference of zero between both sources, suggesting high compliance in the teacher presence data. Panel (b) plots the distribution of the differences for enrolled students. Although there is a high mass around zero, it tends to go toward negative values, suggesting the census data is reporting higher values for students enrolled.

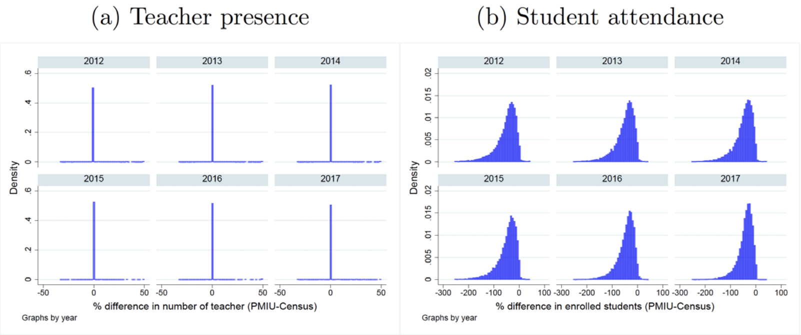
Figure B2: Data Validation: Monthly PMIU vs. Census



(a) Teacher presence (b) Student attendance (c) Functional facilities

Note: This figure plots the distribution of the number of teachers and the number of students enrolled, comparing the monthly data and the annual school census. Both variables are smoothed by a log + 1 transformation to account for the zeros in the data reporting. For the monthly data, only October is chosen, as it is the month in which the census was also collected. The functional facilities variables are measured as 1 if the school reported that the infrastructure was functional, and 0 otherwise. Panel (c) then plots the percentage of coincidence in the reporting between the two sources.

Figure B3: Data Validation: Distributions of the Differences in Reporting Between Monthly (PMIU) vs. Census



(a) Teacher presence (b) Student attendance

Note: This figure plots the distribution of the differences in the reporting number of teachers and number of students enrolled between both sources as percentage change (PMIU - Census)/PMIU. The differences are obtained by comparing the values reported for each school from both sources. For the monthly data, only October is chosen as it is the month in which the census data was also collected. The figure drops for each variable in the data below percentile 1 and above percentile 99.

Creating a ranking of district officer positions

District coordination officers (DCOs) have a supervisory role for all public-sector service delivery at the district level, including education delivery. They can, therefore, be rewarded or punished in terms of transfers to more preferred or less preferred postings, based on educational outcomes at the district level. In order to estimate the effect of the accountability system on the career trajectory of DCOs, we collected information about the postings12 of past DCOs for each district. To ascertain whether a DCO was rewarded or punished, we ranked all designations by seniority and used the ranking to determine whether each position represented a promotion or a demotion. The ranking of designations was generated through extensive research about seniority levels within the Pakistani bureaucracy and was vetted by two senior bureaucrats.

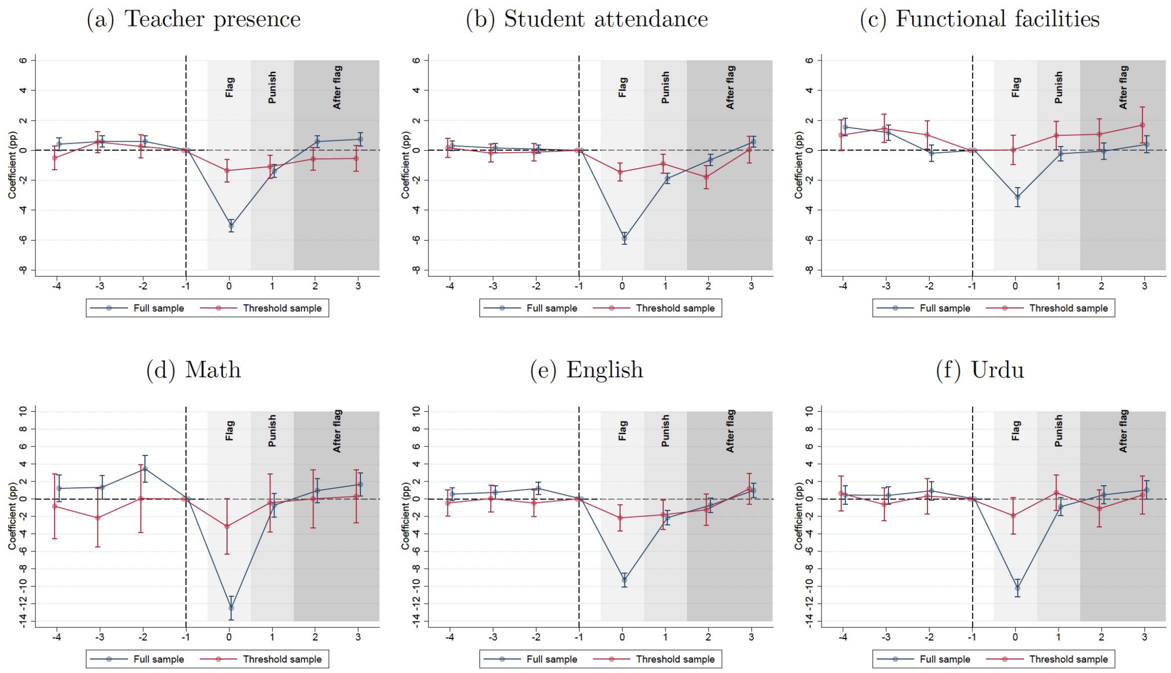
[12] At a certain point in time.

C. Appendix: Additional results

Robustness of the stacked design

To further test the sensitivity of the stacked structure, we first plotted the event study from estimating equation 1 using a lower number of periods after the punishment phase. Figure C1 reports the results, with the y-axis reporting the coefficient β from the equation in percentage points. It can be observed that the coefficients before the flagging are non-significant, and the trends after the flagging suggest reversion to the mean as in the extended event study (Figure 6).

Figure B2: Data Validation: Monthly PMIU vs. Census


(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

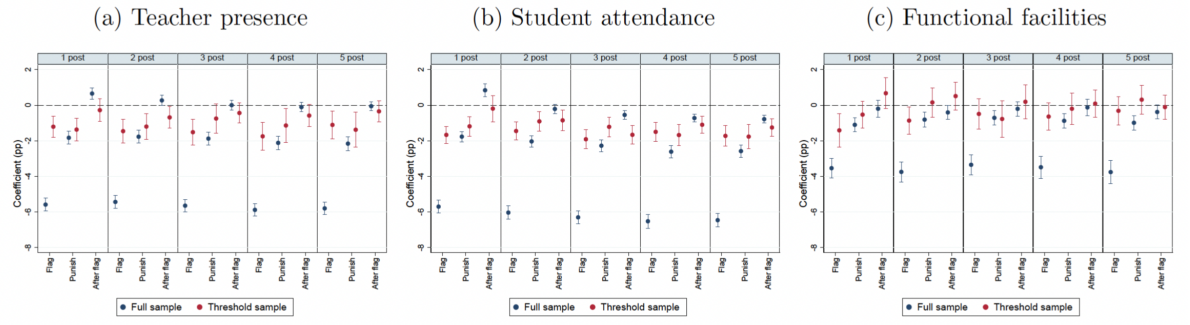
(e) English

(f) Urdu

Note: This figure displays results from estimating an event study based on equation 1, using -1 as the base period and only three periods after the flagging. The blue line accounts for the results on the full sample, while the red accounts for the results using the threshold sample. Flagging is based on the same variable that the one its been observed. Error bars at 95 percent are presented for each coefficient

which clearly shows the immediate version to the mean after the negative shock. Finally, the AfterFlag coefficients remain close to zero. Furthermore, we note that for the case of student attendance, where there are significant and negative results, they are still close to zero and capture some persistence in the recovery, which still suggests the lack of effects of the monitoring system in improving outcomes.

Figure C2: Average Effects by Additional After Flag t - Flagging Effect on Performance


(a) Teacher presence

(b) Student attendance

(c) Functional facilities

Note: This figure displays results from estimating the main specification, adding from one to five additional post periods in the stacked design. Each panel plots for a particular additional t the Flag, Punish and Afterflag effect. Error bars at the 95 percent level are presented for each coefficient.

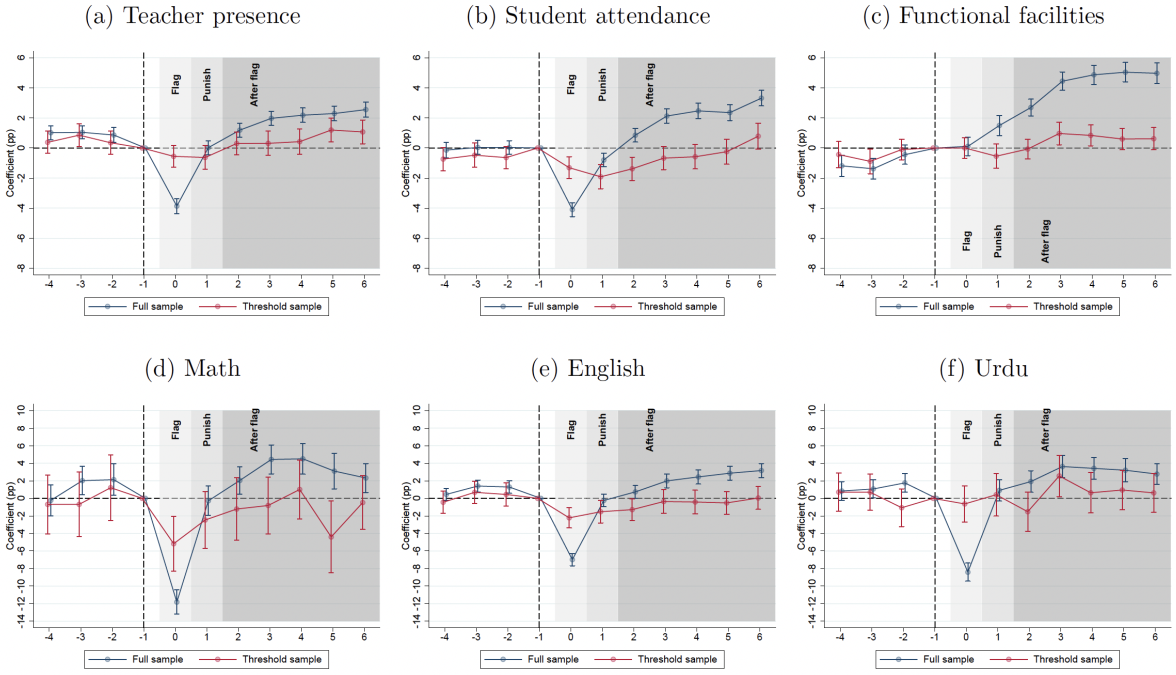
Additional considerations from the monitoring system

In this section we test the robustness of the results under alternative considerations of the flagging system and under alternative estimations. Figure C3 reports the event study from estimating equation 1 using flagging history fixed effects, which compares the maraakiz that had the same path of flagging in the periods before negative shock. Because flagging history fixed effects is not a markaz attribute, the term Tmde from equation 1 is not absorbed, so the positive effects of the interactions should be compared against the Tmde coefficient (hence the positive effects of the event study are overestimated). Table C1 reports the average results of the specification with the modified fixed effects. Although the full sample reports positive coefficients, we note that the average effects of Tmde are negative and higher in absolute magnitude. Furthermore, the positive effects observed are reduced by the negative shock. For the threshold sample, there are no significant effects after the flagging, even without accounting for Tmde, suggesting the lack of an effect of being flagged.

An alternative consideration of analysis consist in the revision of a different threshold for the treatment definition. Although the monitoring system focused more in the red flagging, it is still defined under an arbitrary threshold, so there might be other margins by which effects of being flagged might be observed. We test whether being orange flagged captures some effects of being flagged non-observed under the red flag. Figure C4 plot the coefficients of the event study of the effect of orange flagging on performance. The results suggest that there are not significant effects after the recovery from the negative shock.

Alternatively, although the maraakiz are the main unit of treatment, as they have specific public officials who are in charge to be monitored, there was also reporting of flagging at the tehsil-wing level. Figure C5 plots the results from estimating equation 1 for the tehsil-wing treatment level, including tehsil-wing fixed effects and standard errors clustered at the same level of the treatment. The results suggest a non-significant impact of tehsil-wing flagging on performance; nevertheless, the estimates are more noisy, particularly for the scores variables.

Figure C2: Average Effects by Additional After Flag t - Flagging Effect on Performance


(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

(e) English

(f) Urdu

Note: This figure presents the results from estimating an event study based on equation 1, using -1 as the base period and using flagging history fixed effects. The blue line accounts for the results on the full sample, while the red accounts for the results using the threshold sample. Flagging is based on the same variable that the one its been observed. Error bars at the 95 percent level are presented for each coefficient.

We also tested the robustness of the results considering changes in the datapack structure. Although the instrument for monthly data collection was intended to report school outcomes, it has undergone changes in the amount of information it reports. After December 2015 and January 2017, the datapacks started to include information on more variables, which, in the end, might have affected the response to the availability of the information. We present in Table C2 the average results from the estimation of equation 1, separating the sample based on the datapacks structure. Overall, the results for all datapacks remain similar, as the effects on the after flag period are always closer to zero or non-significant.

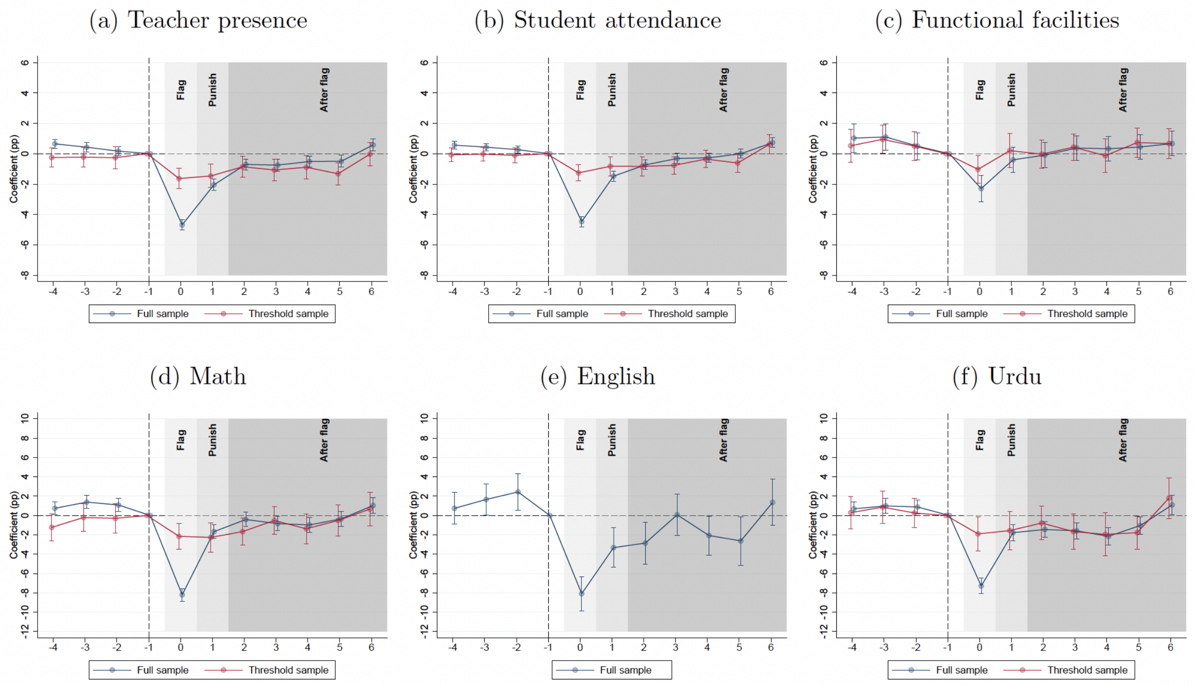
Finally, Table C3 presents the results from the heterogeneity analysis of the coincidence between flagging and district meetings, from a modified version of equation 1, including the triple interaction between being flagged and being in a top/bottom district after the flagging.

Table C1: Monitoring Effect on Performance: Markaz flagging - Flagging History Fixed Effects

Panel A: Outcomes						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T	-2.31***	-0.86***	-3.70***	-1.18***	-8.97***	-1.90***
	(0.11)	(0.16)	(0.16)	(0.18)	(0.29)	(0.19)
T × Flag	-4.64***	-0.97***	-4.08***	-0.84***	0.94***	0.38
	(0.20)	(0.28)	(0.20)	(0.27)	(0.25)	(0.28)
T × Punish	-0.76***	-1.05***	-0.79***	-1.45***	2.34***	-0.17
	(0.19)	(0.33)	(0.19)	(0.30)	(0.28)	(0.35)
T × After flag	1.22***	0.23	2.18***	0.051	5.23***	0.95***
	(0.13)	(0.20)	(0.18)	(0.23)	(0.24)	(0.24)
N. of obs.	1,853,281	325,826	1,884,071	431,981	1,880,251	267,131
Mean of Dep. Var. before	91.0	86.9	87.4	86.2	93.0	91.1
R^2	0.040	0.031	0.17	0.096	0.27	0.12
Panel B: Scores						
Dependent variable:	Math		English		Urdu	
T	-3.81***	0.13	-3.43***	-1.04***	-3.74***	-0.81*
	(0.42)	(0.67)	(0.17)	(0.27)	(0.28)	(0.44)
T × Flag	-12.8***	-5.04***	-7.79***	-2.36***	-9.36***	-0.80
	(0.57)	(1.19)	(0.27)	(0.43)	(0.41)	(0.78)
T × Punish	-1.21*	-2.32*	-1.04***	-1.67***	-0.021	0.26
	(0.67)	(1.22)	(0.29)	(0.49)	(0.51)	(0.95)
T × After flag	2.15***	-0.98	1.46***	-0.62*	1.95***	0.34
	(0.51)	(0.85)	(0.22)	(0.34)	(0.34)	(0.57)
N. of obs.	518,137	22,472	477,348	158,214	508,398	37,666
Mean of Dep. Var. before	86.1	71.4	73.2	69.7	83.2	70.9
R^2	0.078	0.14	0.062	0.055	0.087	0.12
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Flag history FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

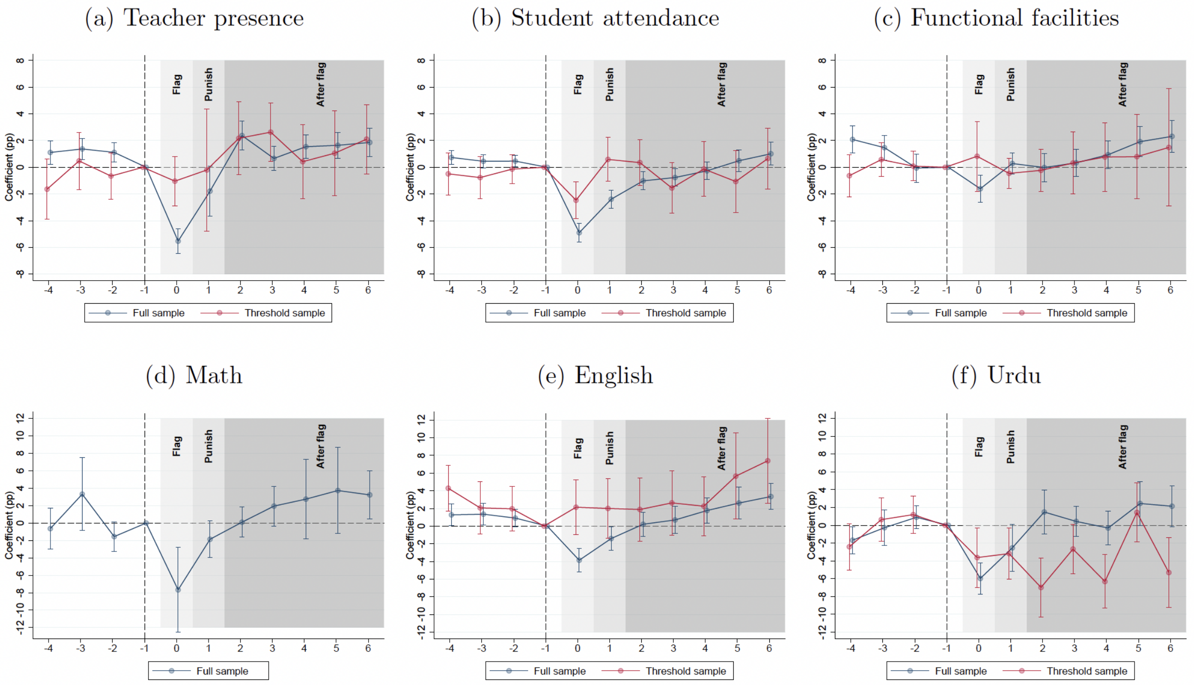
Note: T is equal to 1 for the flagged maraakiz. Flag is equal to 1 for the period in which the information was collected and when the AEOs were flagged. Punish is equal to 1 for the period in which the reports were distributed and the meeting with the punishment happened. After flag is equal to 1 for the periods after the meeting. The threshold sample accounts for the schools in a markaz that lie within the bandwidth obtained through RD methods. Standard errors, clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure C4: Event Study: Flagging Effect on Performance - Orange Threshold



Note: This figure presents the results from estimating an event study based on equation 1, using -1 as the base period and using as treatment those maraakiz that lie below the orange threshold of each variable. The blue line accounts for the results on the full sample, while the red accounts for the results using the threshold sample. Flagging is based on the same variable as the one that has been observed. if there were not enough observations to build a threshold sample for English scores based on the orange threshold, we reported only the full sample for the variable. Error bars at the 95 percent level are presented for each coefficient.

Figure C5: Event Study: Flagging Effect on Performance - Tehsil-Wing Flagging



Note: This figure presents the results from estimating an event study based on equation 1, using -1 as base period. The treatment level is the tehsil-wing instead of the markaz, and the regression includes tehsil-wing fixed effects and standard errors clustered at the tehsil-wing level. The blue line accounts for the results on the full sample, while the red accounts for the results using the threshold sample. Flagging is based on the same variable as the one that has been observed. If there were not enough observations to build a threshold sample for math scores based on the tehsil-wing flagging, we reported only the full sample for the variable. Error bars at the 95 percent level are presented for each coefficient.

Table C2: Monitoring Effect on Performance by Datapack: Markaz Flagging

Panel A: Datapack 1						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T × Flag	-4.75***	-1.02**	-6.19***	-1.56***	-3.73***	-0.31
	(0.20)	(0.47)	(0.24)	(0.35)	(0.37)	(0.44)
T × Punish	-1.66***	-1.60***	-3.03***	-2.03***	-0.94***	0.28
	(0.21)	(0.51)	(0.22)	(0.44)	(0.23)	(0.45)
T × After flag	-0.32**	-0.79**	-0.99***	-1.53***	-0.35*	-0.12
	(0.16)	(0.37)	(0.14)	(0.28)	(0.21)	(0.39)
N. of obs.	1,209,089	245,018	1,041,900	213,132	1,064,203	120,743
Mean of Dep. Var. before	89.7	86.3	85.3	86.5	91.7	92.4
R^2	0.053	0.038	0.22	0.099	0.37	0.24
Panel B: Datapack 2						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T × Flag	-8.70***	-1.36*	-7.55***	-2.23***	-3.39***	0.065
	(0.32)	(0.75)	(0.29)	(0.60)	(0.45)	(1.23)
T × Punish	-3.43***	-0.69	-1.10***	-1.26**	-0.77	1.15
	(0.49)	(1.23)	(0.26)	(0.55)	(0.52)	(0.82)
T × After flag	0.73***	0.68	-0.29*	-0.53*	-0.70	-0.64
	(0.23)	(0.57)	(0.16)	(0.30)	(0.65)	(1.17)
N. of obs.	413,719	36,056	398,525	31,151	408,683	7,893
Mean of Dep. Var. before	93.2	86.8	91.0	84.8	98.3	91.7
R^2	0.071	0.082	0.12	0.12	0.084	0.060
Panel C: Datapack 3						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T × Flag	-11.3***	1.74	-6.39***	-3.08**	-6.56***	-2.23
	(1.79)	(5.44)	(0.97)	(1.38)	(1.34)	(2.46)
T × Punish	-8.73***	-2.71	-2.05***	0.030	-2.84***	-1.32
	(2.04)	(4.57)	(0.60)	(1.33)	(0.96)	(1.53)
T × After flag	-3.45***	4.76*	-0.59	-1.27	-0.25	1.27
	(1.02)	(2.66)	(0.37)	(0.79)	(0.70)	(1.25)
N. of obs.	64,419	1,071	86,407	4,737	81,605	2,503
Mean of Dep. Var. before	95.5	90.9	91.8	85.9	98.3	91.9
R^2	0.077	0.088	0.13	0.14	0.086	0.095
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Note: Here T is equal to 1 for the flagged maraakiz. Flag is equal to 1 for the period in which the information was collected and when AEOs were flagged. Punish is equal to 1 for the period in which the reports were distributed and the meeting with the punishment happened. After flag is equal to 1 for periods after the meeting. The threshold sample accounts for the schools in maraakiz that lie in within the bandwidth obtained through RD methods. Standard errors, clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C3: Monitoring Effect on Performance: District Ranking and Markaz Flagging

Panel A: Bottom districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Bottom × Flag × Meeting	-0.15	-0.68	-0.045	0.67	-0.51	-0.63
	(0.51)	(0.61)	(1.23)	(1.27)	(0.72)	(0.83)
Bottom × Meeting	0.55**	0.61**	0.99	1.12*	0.29	0.35
	(0.26)	(0.29)	(0.86)	(0.60)	(0.40)	(0.56)
Flag × Meeting	-4.86***	-4.03***	-4.68***	-5.26***	-0.84	-0.47
	(0.19)	(0.50)	(0.59)	(1.00)	(0.59)	(0.52)
Bottom × Flag × After meeting	0.83	0.36	-1.26	-0.53	-0.55	-0.20
	(0.52)	(0.65)	(0.98)	(1.05)	(0.69)	(0.73)
Bottom × After meeting	0.22	0.18	1.12*	1.52**	0.61	0.31
	(0.30)	(0.36)	(0.59)	(0.57)	(0.55)	(0.51)
Flag × After meeting	-0.34*	0.11	0.81***	0.40	1.28***	1.09**
	(0.19)	(0.47)	(0.18)	(0.59)	(0.26)	(0.47)
N. of obs.	3,063,826	583,327	3,063,401	583,158	3,009,824	565,830
Mean of Dep. Var. before	91.4	90.1	88.8	86.0	92.5	90.0
R^2	0.028	0.033	0.13	0.16	0.17	0.20

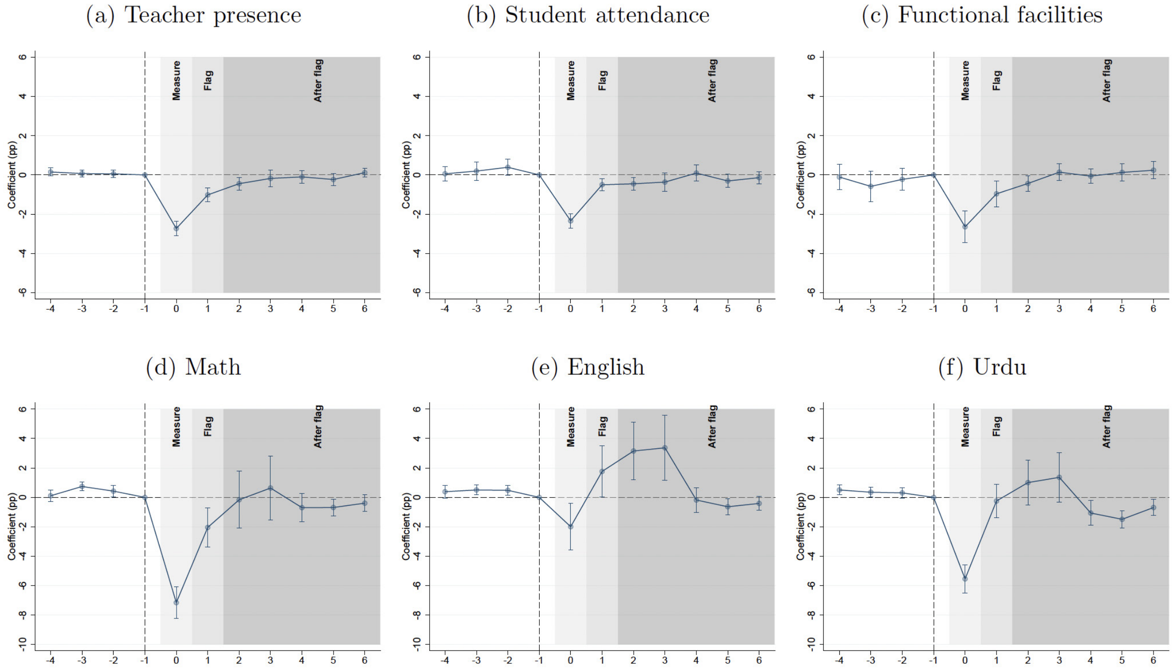
Panel B: Top districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Top × Flag × Meeting	0.40	0.25	-0.48	-3.34*	1.24	3.27
	(0.53)	(0.89)	(1.56)	(1.92)	(0.91)	(3.19)
Top × Meeting	0.48	0.31	-1.28**	-0.64	0.0020	-0.50
	(0.30)	(0.30)	(0.56)	(0.60)	(0.33)	(0.85)
Flag × Meeting	-4.74***	-5.18***	-5.02***	-2.71***	-0.78	-3.32
	(0.22)	(0.64)	(0.54)	(0.90)	(0.56)	(3.27)
Top × Flag × After meeting	-0.52	-1.39	1.12	-0.73	1.13	0.91
	(0.54)	(1.04)	(0.90)	(1.06)	(1.27)	(0.92)
Top × After meeting	0.81***	0.94***	-0.12	-0.32	-0.11	0.49
	(0.24)	(0.22)	(0.43)	(0.60)	(0.19)	(0.39)
Flag × After meeting	-0.030	0.50	0.45**	1.38**	1.28***	1.16**
	(0.14)	(0.69)	(0.18)	(0.67)	(0.26)	(0.52)
N. of obs.	3,111,411	682,498	3,110,817	682,406	3,036,323	672,814
Mean of Dep. Var. before	91.0	91.9	87.4	90.0	91.6	92.7
R^2	0.029	0.028	0.13	0.13	0.17	0.18
Sample	Full	Threshold	Full	Threshold	Full	Threshold
District FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: Here Bottom is equal to 1 for schools in the bottom five districts on the date of a quarterly meeting. The top equals 1 for schools in the top five districts on the date of the quarterly meeting. Flag takes the value of 1 for schools in maraakiz in which the respective outcome was below the threshold on the date of the quarterly meeting. Meeting is equal to 1 for the period in which the quarterly meeting took place. The threshold sample accounts for the schools in the five districts following the top/bottom. Standard errors, clustered by district, are in parentheses. *p < 0.10, ** p < 0.05, *** p < 0.01.

Additional difference-in-difference estimators

We present additional event study specifications to account for alternative difference-in- difference estimators for our main specifications. Figure C6 estimates the event study fol- lowing Sun and Abraham (2021) on the original (non-stacked) data set, under the assumption of a staggered treatment timing, where the flagged maraakiz remain treated after their first occurrence. We show that under a two-way fixed effects dynamic specification with stag- gered treatment adoption, leads and lag coefficients are contaminated by the effect on other relative periods.13 Because the stacked design re-centers the treatment, the adoption timing is not an issue, so coefficients are obtained through two-way fixed effects. Hence, allowing for differential treatment timing might affect the results. Still, the results for each outcome suggest that parallel trends hold, and all follow the trend of reporting a dip, after which the outcome recovers to pre-shock levels, and the effect of the high-frequency oversight scheme is null.

Figure C6: Alternative Difference-in-Differences Specifications - Sun and Abraham (2021)



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

(e) English

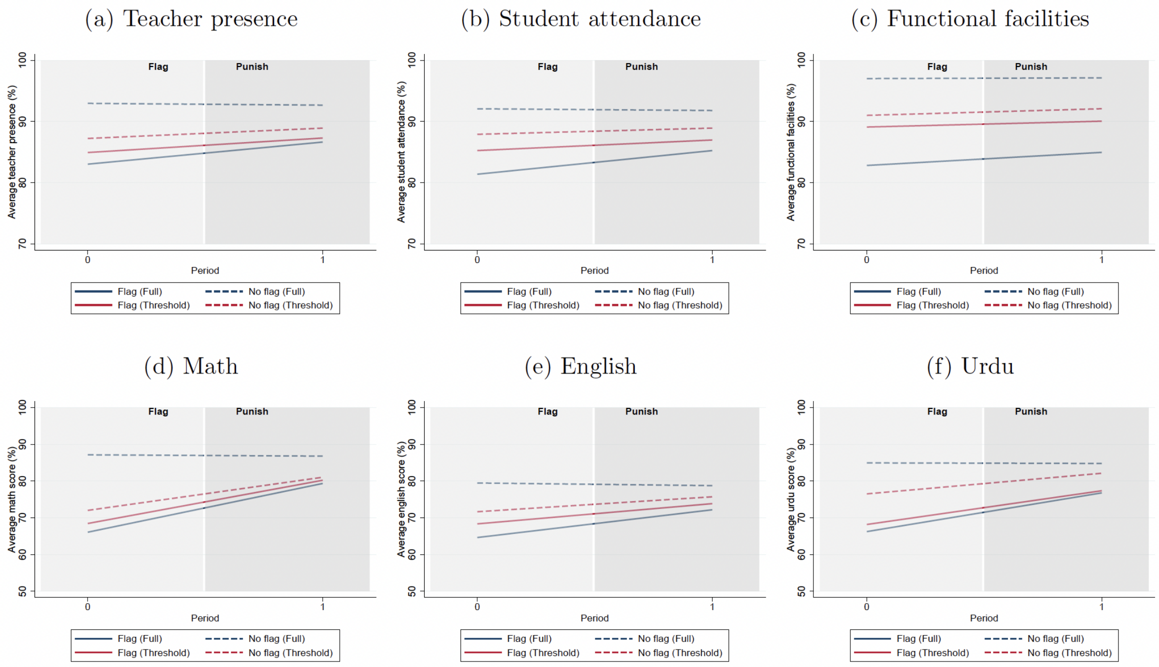
(f) Urdu

Note: This figure presents the results from estimating an event study based on the Sun and Abraham (2021) difference-in-differences estimator, using -1 as the base period. Flagging is based on the same variable as the one that has been observed. Error bars at the 95 percent level are presented for each coefficient.

[13] Following Baker et al. (2022), the estimator of Sun and Abraham (2021) is an special case of the estimator of Callaway and Sant'Anna (2021), when there are no covariates.

Counterfactual analysis and the persistence of public policy

Figure C7: Pre/Post Evolution of Outcomes



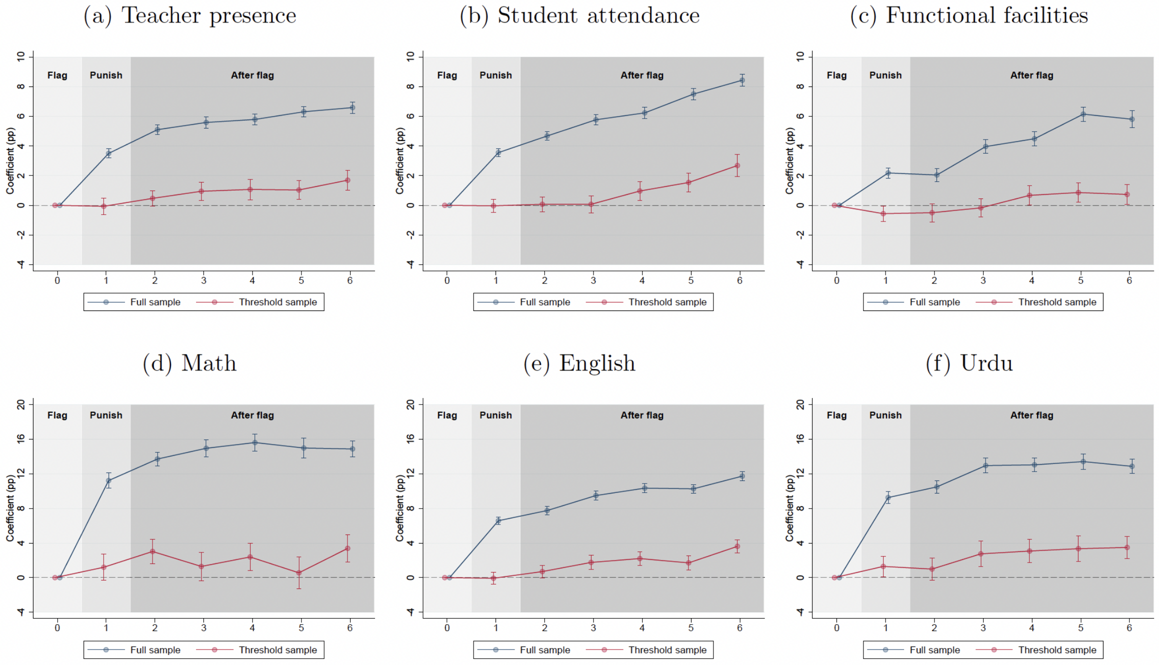
Note: This figure presents the average evolution of schools with flagged (continuous line) and non-flagged (dashed line) AEOs. Flagging is based on the same variable as the one that has been observed. Blue lines represent the evolution for the full sample, while red accounts for the evolution in the threshold sample. The dark dashed line marks -1 as the first relative time. Flag is the period in which the information was collected and when AEOs were flagged. Punish is the period in which the reports were distributed and the meeting with the punishment happened.

Table C4: Immediate Effect on Performance: Markaz Flagging

Panel A: Outcomes						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T	-6.62***	-2.08***	-6.72***	-1.92***	-6.76***	-1.32***
	(0.10)	(0.11)	(0.091)	(0.084)	(0.16)	(0.13)
T × Punish	4.38***	1.05***	5.00***	1.13***	2.57***	0.035
	(0.12)	(0.16)	(0.11)	(0.13)	(0.14)	(0.20)
N. of obs.	1,945,025	252,922	1,836,939	318,436	1,743,559	154,071
Mean of Dep. Var. before	92.5	87.4	90.1	86.6	96.1	91.4
R^2	0.042	0.023	0.17	0.075	0.15	0.030
Panel B: Scores						
Dependent variable:	Math		English		Urdu	
T	-15.8***	-3.23***	-11.1***	-3.15***	-12.8***	-3.76***
	(0.25)	(0.27)	(0.10)	(0.098)	(0.19)	(0.19)
T × Punish	12.3***	2.43***	7.98***	2.21***	9.17***	2.42***
	(0.32)	(0.46)	(0.15)	(0.20)	(0.25)	(0.33)
N. of obs.	835,135	31,966	760,869	166,528	830,410	80,508
Mean of Dep. Var. before	86.8	72.0	76.7	70.8	84.4	75.3
R^2	0.089	0.092	0.094	0.040	0.10	0.13
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Flag history FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Note: Here T is equal to 1 for the flagged maraakiz. We note that in this case T happens also in the Flag period of the previous estimations. Punish is equal to 1 for the period in which the reports were distributed and the meeting with the punishment happened. Flag History is a categorical variable that concatenates 1 or 0 if the maraakiz were flagged in the three periods before. Hence, it groups all the maraakiz that follow the same flagging path, except for the last. Standard errors, clustered by markaz, are in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Figure C8: Event Study: Flagging Effect on Performance - Flagging History Fixed Effects No Pre-Periods)



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

(e) English

(f) Urdu

Note: This figure presents the results from estimating an event study based on equation 1, with no pre- periods. We use 0 as the base period and we use flagging history fixed effects. The blue line accounts for the results on the full sample, while the red accounts for the results using the threshold sample. Flagging is based on the same variable as the one that has been observed.

Impacts on the machinery of the government

Table C5: Monitoring Effect on Other Outcomes: Effort as Mechanism

Flagging	Teacher presence		Student attendance		Functional facilities		Math		English		Urdu	
T × Flag	-0.0035	-0.021	-0.0030	0.0061	0.033**	0.029	0.029***	0.028	0.012**	-0.0057	0.035***	0.014
	(0.0058)	(0.015)	(0.0046)	(0.0087)	(0.014)	(0.030)	(0.0075)	(0.020)	(0.0047)	(0.013)	(0.0080)	(0.019)
T × Punish	0.0013	-0.049**	-0.0035	0.0036	0.036***	0.019	-0.028**	0.025	0.0026	-0.0055	-0.0073	-0.021
	(0.0063)	(0.020)	(0.0050)	(0.011)	(0.010)	(0.027)	(0.011)	(0.022)	(0.0054)	(0.016)	(0.0091)	(0.024)
T × After flag	-0.00072	-0.026	0.0095**	0.021*	0.031***	0.039	0.0020	0.033*	0.0012	0.013	0.0091	0.0099
	(0.0057)	(0.018)	(0.0042)	(0.013)	(0.0091)	(0.025)	(0.0074)	(0.019)	(0.0045)	(0.013)	(0.0074)	(0.019)
N. of obs.	1,059,497	228,333	1,040,426	194,387	924,381	91,743	293,027	18,357	219,337	59,280	295,098	32,612
Mean of Dep. Var. before	0.90	0.81	0.88	0.95	0.87	0.83	0.96	0.95	0.97	0.97	0.96	0.97
R^2	0.33	0.33	0.34	0.24	0.35	0.37	0.22	0.25	0.22	0.26	0.22	0.26
Sample	Full	Threshold	Full	Threshold	Full	Threshold	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: Here visits to schools is measured as a dummy that takes the value of 1 if the schools received a visit by an AEO. Is interpreted as the probability of receiving a visit. T is equal to 1 for flagged maraakiz. Flag is equal to 1 for the period in which the information was collected and when AEOs were flagged. Punish is equal to 1 for the period in which the reports were distributed and the meeting with the punishment happened. After flag is equal to 1 for the periods after the meeting. The threshold sample accounts for the schools in maraakiz that lie in within the bandwidth obtained through RD methods. Standard errors, clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C5: Monitoring Effect on Other Outcomes: Effort as Mechanism

Flagging	Teacher presence		Student attendance		Functional facilities		Math		English		Urdu	
T × Flag	0.0072**	0.013*	-0.0045	0.00072	0.014**	-0.0010	-0.014	-0.020	-0.00053	0.0042	-0.0041	-0.0087
	(0.0035)	(0.0077)	(0.0039)	(0.0096)	(0.0070)	(0.0097)	(0.010)	(0.027)	(0.0066)	(0.019)	(0.0085)	(0.015)
T × Punish	0.00063	-0.015	0.00055	0.0096	0.010	-0.020***	-0.031***	-0.060**	-0.0093	-0.015	-0.014**	-0.0051
	(0.0032)	(0.010)	(0.0038)	(0.0088)	(0.0062)	(0.0072)	(0.0089)	(0.030)	(0.0064)	(0.020)	(0.0069)	(0.015)
T × After flag	0.0011	-0.0088	-0.0015	-0.0037	0.0056	-0.0026	-0.020***	-0.020	-0.0037	0.0064	-0.012**	0.0100
	(0.0029)	(0.0079)	(0.0031)	(0.0074)	(0.0043)	(0.0075)	(0.0066)	(0.015)	(0.0041)	(0.0093)	(0.0059)	(0.0095)
N. of obs.	1,687,402	282,251	1,527,309	249,150	1,581,638	133,762	510,284	29,037	309,060	73,590	490,316	44,037
Mean of Dep. Var. before	0.060	0.057	0.061	0.055	0.061	0.052	0.060	0.082	0.073	0.085	0.065	0.050
R^2	0.16	0.13	0.15	0.13	0.16	0.10	0.23	0.19	0.20	0.19	0.23	0.17
Sample	Full	Threshold	Full	Threshold	Full	Threshold	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: Here visits to schools is measured as a dummy that takes the value of 1 if the schools received a visit by an AEO. Is interpreted as the probability of receiving a visit. T is equal to 1 for flagged maraakiz. Flag is equal to 1 for the period in which the information was collected and when AEOs were flagged. Punish is equal to 1 for the period in which the reports were distributed and the meeting with the punishment happened. After flag is equal to 1 for the periods after the meeting. The threshold sample accounts for the schools in maraakiz that lie in within the bandwidth obtained through RD methods. Standard errors, clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C7: Monitoring Effect on Yearly School Budget

Panel A: Teacher presence flagging					
Dependent variable (in logs):	Funds	Expenses	Non-salary funds	Non-salary expenses	Joint significance
Num times flagged (previous year)	-0.0020	0.14***	0.023**	0.034	0.00 ***
	(0.0071)	(0.020)	(0.010)	(0.028)	Significant
N. of obs.	92,671	184,054	134,530	143,768	
Mean of Dep. Var	44099.3	35711.6	100427.0	46714.1	
R^2	0.41	0.26	0.28	0.30	
Panel B: Student attendance flagging					
Dependent variable (in logs):	Funds	Expenses	Non-salary funds	Non-salary expenses	Joint significance
Num times flagged (previous year)	0.0042	0.038**	-0.0074	0.024	0.23
	(0.0057)	(0.018)	(0.0059)	(0.023)	Not significant
N. of obs.	92,671	184,054	134,530	143,768	
Mean of Dep. Var	44099.3	35711.6	100427.0	46714.1	
R^2	0.41	0.26	0.28	0.30	
Panel C: Functional facilities flagging					
Dependent variable (in logs):	Funds	Expenses	Non-salary funds	Non-salary expenses	Joint significance
Num times flagged (previous year)	0.0032	0.030*	0.011*	0.0090	0.25
	(0.0039)	(0.016)	(0.0061)	(0.018)	Not significant
N. of obs.	92,671	184,054	134,530	143,768	
Mean of Dep. Var	44099.3	35711.6	100427.0	46714.1	
R^2	0.41	0.26	0.28	0.30	
Panel D: Math scores flagging					
Dependent variable (in logs):	Funds	Expenses	Non-salary funds	Non-salary expenses	Joint significance
Num times flagged (previous year)	-1.95**	0.017	-0.42	0.072	0.61
	(0.85)	(0.13)	(0.37)	(0.20)	Not significant
N. of obs.	92,698	184,076	134,557	143,780	
Mean of Dep. Var	44099.3	35711.6	100427.0	46714.1	
R^2	0.41	0.26	0.28	0.30	
Panel E: English scores flagging					
Dependent variable (in logs):	Funds	Expenses	Non-salary funds	Non-salary expenses	Joint significance
Num times flagged (previous year)	0.11	0.052	-0.20***	-0.0064	0.05**
	(0.19)	(0.039)	(0.059)	(0.042)	Significant
N. of obs.	92,671	184,054	134,530	143,768	
Mean of Dep. Var	44099.3	35711.6	100427.0	46714.1	
R^2	0.41	0.26	0.28	0.30	
Panel F: Urdu scores flagging					
Dependent variable (in logs):	Funds	Expenses	Non-salary funds	Non-salary expenses	Joint significance
Num times flagged (previous year)	-0.17	0.15	-0.13	0.045	0.69
	(0.49)	(0.12)	(0.17)	(0.13)	Not significant
N. of obs.	92,671	184,054	134,530	143,768	
Mean of Dep. Var	44099.3	35711.6	100427.0	46714.1	
R^2	0.41	0.26	0.28	0.30	
Markaz FE	Yes	Yes	Yes	Yes	
Time FE	Yes	Yes	Yes	Yes	
District time trends	Yes	Yes	Yes	Yes	

Note: The dependent variable measures are in logs. The explanatory variable is the number of times a school was flagged in a year. The regression includes markaz fixed effects, time fixed effects, and district time trends. Standard errors are clustered at the level. Joint significance reports the result from a Wald test of joint significance between the coefficients of the regressions for the four dependent variables for each outcome. Specifically, it reports the level of significance at which the null hypothesis of all coefficients being jointly equal to zero can be rejected. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C8: Monitoring Effect on Labor Markets: District Ranking on Change of District Officers

Panel A: Bottom districts		
Dependent variable:	Change of DC	
Bottom × Meeting	0.028	0.028
	[0.042]	[0.057]
Bottom × After meeting	0.019	0.015
	[0.027]	[0.051]
N. of obs.	2,921	605
Mean of Dep. Var. before	0.060	0.066
R^2	0.31	0.46
Panel B: Top districts		
Dependent variable:	Change of DC	
Top × Meeting	0.021	0.040
	[0.062]	[0.061]
Top × After meeting	-0.041	-0.044
	[0.029]	[0.041]
N. of obs.	3,025	685
Mean of Dep. Var. before	0.053	0.047
R^2	0.28	0.32
Sample	Full	Threshold
District FE	Yes	Yes
Time FE	Yes	Yes

Note: Here Bottom is equal to 1 for schools in the bottom five districts on the date of a quarterly meeting. Top equals 1 for schools in the top five districts on the date of the quarterly meeting. Meeting is equal to 1 for the period in which the quarterly meeting takes place. Threshold samples account for the schools in the five districts following the top/bottom. The data is aggregated at the district/date level as the outcomes do not vary by school within a district. Bootstrapped standard errors are in brackets. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table C9: Monitoring Effect on Labor Markets: Current Position of District Officers

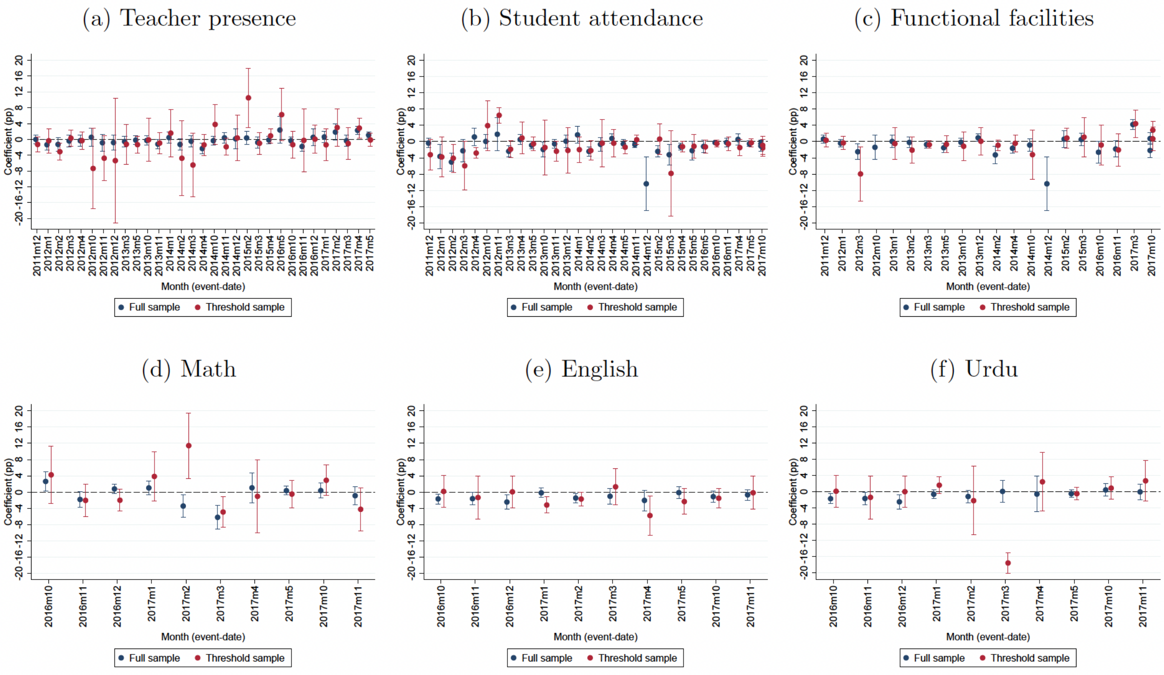
Panel A: Bottom districts	
Dependent variable:	Rank of current employment
Months in bottom district	-1.02
	[1.19]
N. of obs.	96
Mean of Dep. Var. before	3.08
R^2	0.0045
Panel B: Top districts	
Dependent variable:	Rank of current employment
Months in Top district	1.27
	[1.50]
N. of obs.	93
Mean of Dep. Var. before	2.67
R^2	0.010

Note: Bootstraped standard errors are in brackets. * p < 0.10, ** p < 0.05, *** p < 0.01.

Other robustness checks

We also tested additional mechanisms by which results might be confounded. First, we estimated equation 1 for each specific event panel to check for the robustness of the results to time shocks. Figure C9 reports the coefficients for both samples and each outcome in the Afterflag period. Overall, in most of the event panels, there appear to be non-significant re- sults, which supports the evidence that, on average, centralized monitoring has not improved schools' performance.

Figure C9: Seasonality: Monthly Effects of Flagging



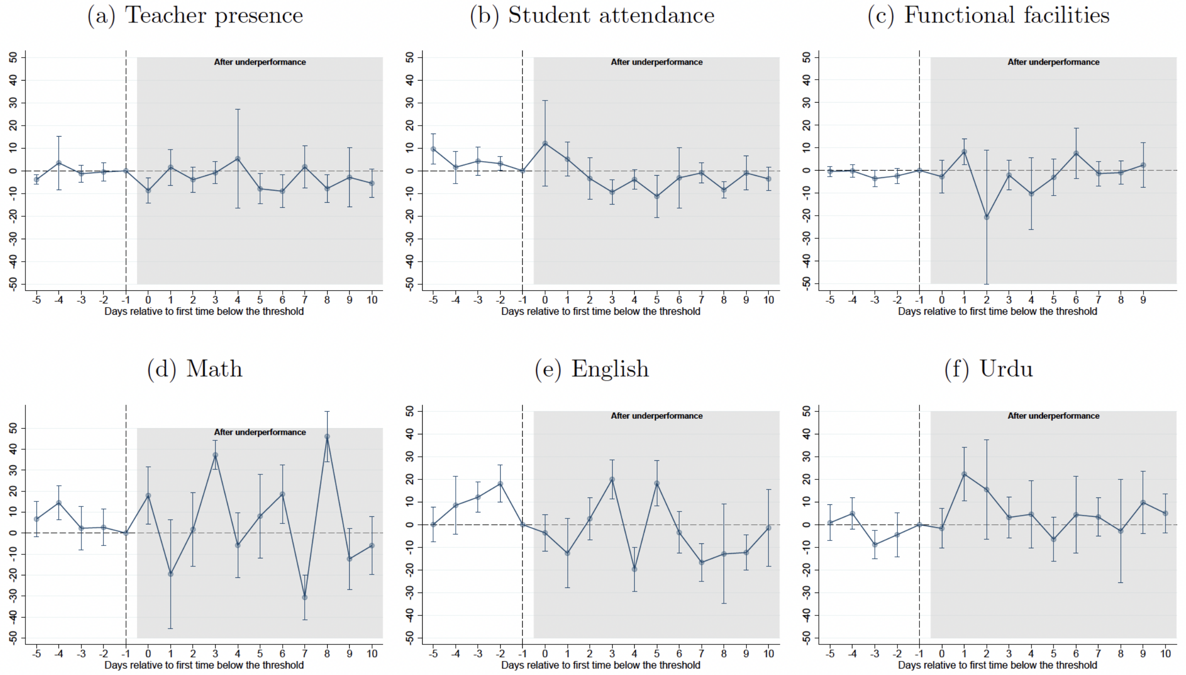
(a) Teacher presence (b) Student attendance (c) Functional facilities

(d) Math (e) English (f) Urdu

Note: This figure presents the results from estimating for each event time the effects for the AfterFlag period, which captures the effect of flagging after the reversion to the mean. Error bars at the 95 percent level are presented for each coefficient.

Finally, we tested the reversion to the mean hypothesis by identifying whether there exists anticipation to the flagging. The premise follows the assumption that a markaz might start recovering before receiving the flagging if the AEO knows they might be flagged at the end of the month. To test for this, we estimated a daily event study in which treatment starts once the average outcome of the visited schools on a particular day lies below the flagging threshold. In such a case, we assume that AEOs might identify the potential flagging and hence react in the days afterward. The results in Figure C10 suggest that no reaction exists in response to being below the threshold for the first time in the month.

Figure C10: Flagging Anticipation

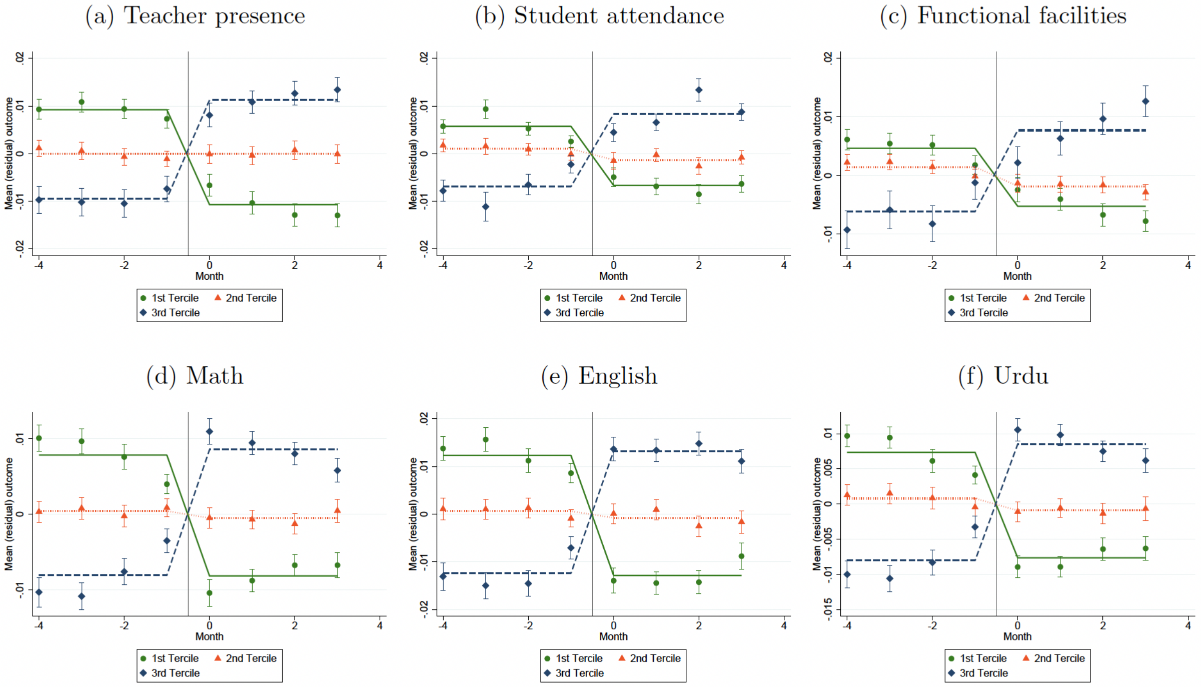

(a) Teacher presence (b) Student attendance (c) Functional facilities

(d) Math (e) English (f) Urdu

Note: This figure presents the results from estimating an event study for the daily average of the outcomes in the month of a flagging. The base period consists in the day just before the average of the visited schools in a lie below the threshold. Error bars at the 95 percent level are presented for each coefficient.

Additional results - Modeling effective centralized labor market management

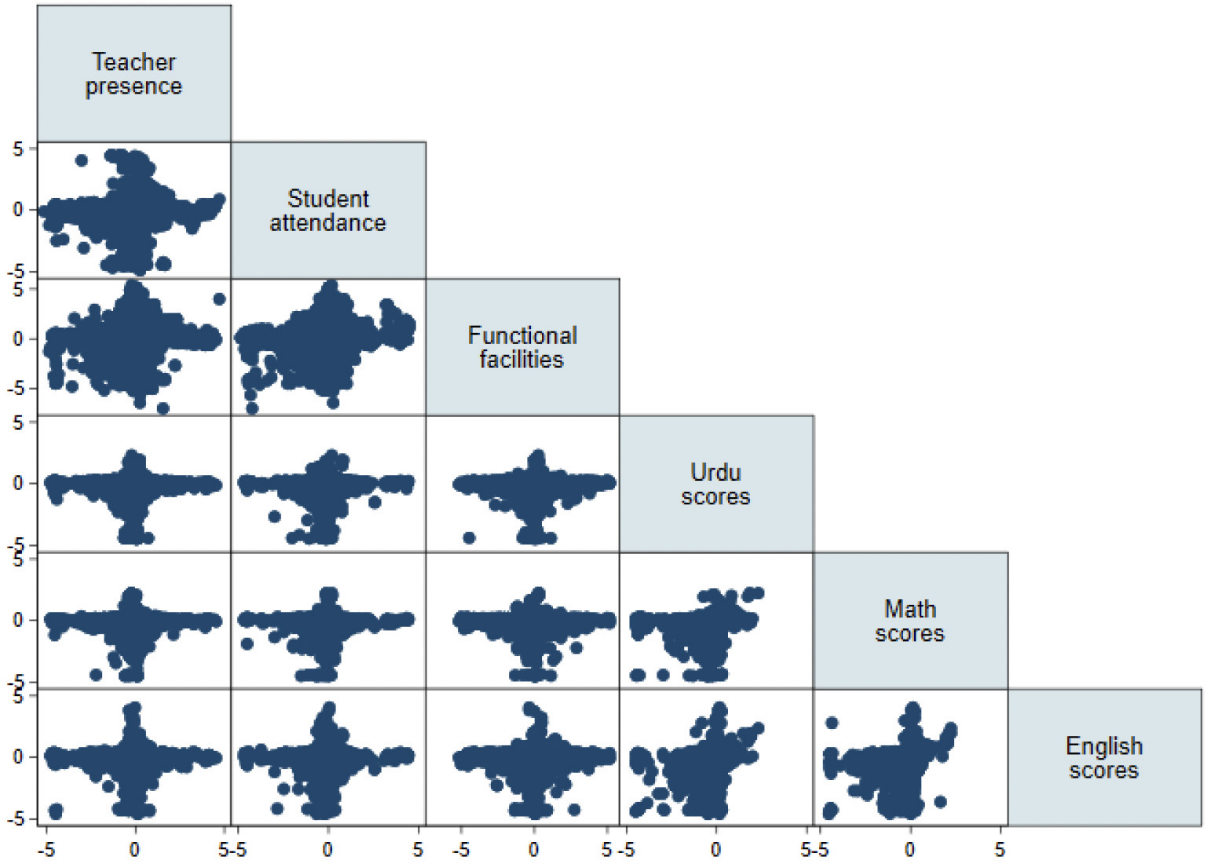
Figure C11: Testing for Sorting by Drift Component


(a) Teacher presence (b) Student attendance (c) Functional facilities

(d) Math (e) English (f) Urdu

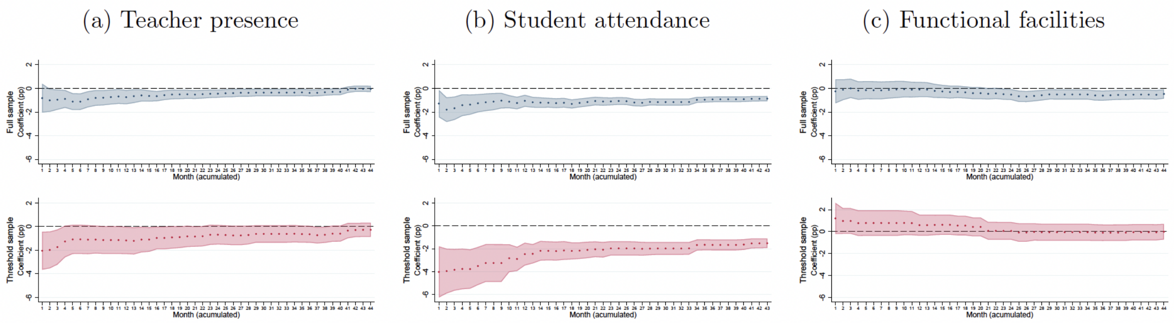
Note: This figure plots the mean (trend-adjusted) log outcomes and the 95 percent confidence intervals relative to change of head teacher, classified by tertiles of change in quality.

Figure C12: Correlation Between Different Outcome-Based Head Teacher Quality Measures



Note: This figure presents the correlation between each head teacher quality measure, obtained through estimating equation 3 on each of the log outcomes.

Figure C13: Accumulated Flagging Effects by Month - After Flag Accumulated Effect


(a) Teacher presence

(b) Student attendance

(c) Functional facilities

Note: This figure presents the results from estimating the main specification, accumulating one month at a time. Standard errors are clustered at the level.

List of Acronyms

AEO	assistant education officer
CEO	chief executive officer (education)
CM	chief minister
DCO	district coordination officer
DDEO	deputy district education officer
DEA	district education authority
DEO	district education officer
DMO	district monitoring officer
EE-F	elementary education female
EE-M	elementary education male
MEA	monitoring and evaluation assistant
PMIU	Program Monitoring and Implementation Unit
SE	secondary education