Command and Can't Control: Assessing Centralized Accountability in the Public Sector*

Saad Gulzar
Juan Felipe Ladino
Muhammad Zia Mehmood
Daniel Rogger†

December 12, 2023

Abstract

A long-established approach to management in government has been the transmission of information up a hierarchy, centralized decision-making by senior management, and corresponding centralized accountability; colloquially known as 'command and control'. This paper examines the effectiveness of a centralized accountability system implemented at scale in Punjab, Pakistan for six years. The scheme automatically identified poorly performing schools and

*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, Michael Callen, Alessandra Fenizia, Koen Geven, Dan Honig, Clare Leaver, Rabea Malik, Imran Rasul and Martin Williams for their guidance and useful comments; and, seminar participants at Berkeley, the Education Commission, Georgetown, the Institute of Development and Economic Alternatives, the Institute for Fiscal Studies, Oxford, and the World Bank. All errors are our own. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments 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.

jurisdictions for the attention of central management. We find that flagging of schools and corresponding de facto punishments had no impact on school or student outcomes. We use detailed data on key elements of the education production function to show that command and control approaches to managing the general public sector do not induce bureaucratic action towards improvements in government performance.

JEL CODES: D73, H11, H83

1 Introduction

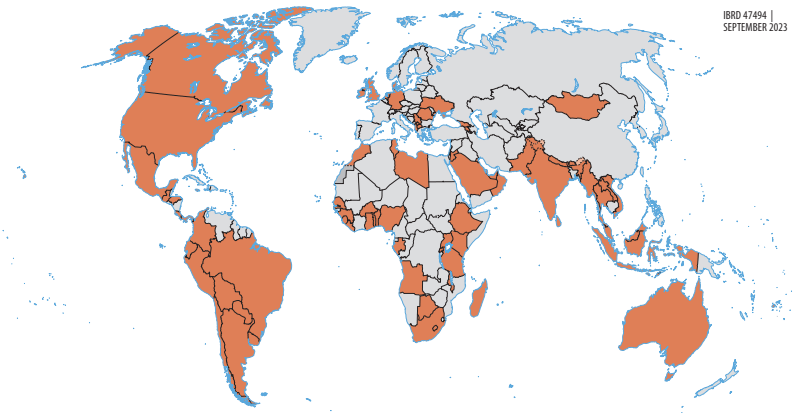
How the bureaucracy performs is fundamental to the provision of high-quality public services in the developing world (Besley et al., 2022). Recent approaches to bolstering the functioning of public administration have focused on de jure improvements in formal contracting environments such as introducing pay-for-performance (Muralidharan and Sundararaman, 2011; Dal Bó, Finan and Rossi, 2013; Ashraf, Bandiera and Jack, 2014; Deserranno, 2019; Leaver et al., 2021). However, the vast majority of reforms to government administration implemented at scale relate to shaping the de facto incentive environment in the bureaucracy instead of introducing changes in legal and fiscal environments. The Global Survey of Public Servants (Schuster et al., 2023), run in 35 countries, reports that only 31% of public servants perceive their public service as actualizing de jure performance incentives, while 76% state that de facto reward systems are in operation.

A canonical de facto bureaucratic reform is *command-and-control* management, or hierarchical systems of control where officials are expected to follow centrally determined directions or face punishment. Finer (1997)'s magisterial overview of administrative arrangements of government throughout history emphasizes the continuous efforts of monarchies and autocracies towards the centralization of information and control around a sovereign. Modern military administrators across the world rely on command-and-control for effective governance across the hierarchy (Wilson, 1989; Hoehn, Campbell and Bowen, 2021).

Faced with constraints on de jure changes in public sector incentives, civilian public sector bureaucracies have been attracted to adopt a command and control model. Fol-

lowing the purported success of British Prime Minister Tony Blair's 'delivery unit',[1] over 80 countries have set up centralized routines and offices (see Figure 1) that "combine functions such as target-setting, monitoring, accountability, and problem-solving with the aim of rapidly improving bureaucratic performance and service delivery" (Education Commission, 2023, p. 7). What distinguishes these reforms is the extraordinary political and executive backing they received around the world. However, evidence on the efficacy of applying command and control approaches to modern public administrations at scale remains scarce.

Figure 1: Countries adopting the command-and-control delivery approach (shaded)



Source: Mansoor et al. (2023)

We study such a scheme implemented at scale in the education public administration of Punjab, Pakistan, where monthly education data from over 50 thousand public schools was channeled to the highest executive authority and used to set targets and establish accountability throughout the organization. This command-and-control scheme in Punjab is considered a showpiece of the centralized accountability delivery model: it was implemented to a very high standard for over six years, was advised by top experts in the world, and had the full backing and involvement of the most senior members of the executive (Barber, 2013; Chaudhry and Tajwar, 2021; Malik and Bari, 2022).[2]

[1] See The History of Government Blog (2022) for more details.

[2] Education Commission (2023) write that "the chief minister... attended all 39 stocktake meetings to hold districts accountable, and took action to solve implementation bottlenecks in the quarterly

Our analysis focuses on the efficacy of the scheme as a driver of improved educational outcomes. We collect the administrative data from all 52,000 schools in Punjab from December 2011 to May 2018 on which the scheme was built and digitize the monthly reports created for senior managers that flagged performing and underperforming school districts.[3] The monitoring reports present performance metrics drawn from this data, aggregated at the administrative unit, for a range of school outcomes including teacher presence, student attendance, functional facilities, and from September 2017, student test scores on standardized exams. Using this data, we examine how senior officials' high-frequency monitoring of public services and efforts to exert control impact subsequent school performance.

To more deeply assess the impacts of command and control approaches on public administration, we also collect data on key elements of the education administration related to financial and personnel resources, bureaucratic attention to individual schools, and the career progressions of affected officials. This data allows us to unpack the impact of the scheme across the hierarchical chain and explore a broad range of bureaucratic responses to the 'command-and-control' system.

In our core specifications, we use a stacked difference-in-differences design (Cengiz et al., 2019; Baker, Larcker and Wang, 2022) to assess the impact of a public official being flagged by the monitoring system on educational outcomes under their responsibility. An official is only flagged if a sufficient percentage of schools within their jurisdiction have fallen below a threshold in the outcome of interest. The first difference in our research design compares schools in flagged and non-flagged administrative units. To make this comparison sharper, we additionally hone in on a sample of administrative units, labeled the 'threshold sample', that lies just above or below the threshold for flagging so that both administrative units see a comparable drop in the outcome of interest, but only one of them is flagged. The second difference

high-stakes meetings" (p.16). A qualitative review of the scheme stated "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 and Bari, 2022). The implementation in Punjab is highlighted as one of the success stories around the world. Reviewing the scheme in an interview in 2017, Michael Barber, one of the architects of the delivery approach around the world, stated, "Punjab is unique ... across the whole world for combining deliverology with really good and modern technology."

[3]The school-level data was collected by an agency within the education sector that is fully independent of the bureaucrats being monitored, and we validate its quality by using a distinct set of independent assessments.

compares the trajectory of treated and non-treated administrative units over time so that we can assess their response to shocks in the outcomes of interest.

We find precisely estimated evidence that the scheme had no substantive impact on targeted school outcomes: teacher and student attendance, functional school facilities, as well as English, Mathematics, and Urdu test scores. Though there is a small increase in the rate at which teachers return to schools after an absence, the limited magnitude on an outcome clearly within the authority of public managers - a 2 percentage point faster month-on-month improvement - in fact underlines the limitations of the scheme.

Despite these broadly null effects the program was maintained, and further developed, for 6 years. A potential reason for the persistence of the program is that a naive examination of before-after comparisons yields a strong positive effect of the program. Many outcomes in the policy domain exhibit reversion to the mean following idiosyncratic shocks, such as student test scores (Chay, McEwan and Urquiola, 2005). Our paper extends this finding to the overarching machinery of public administration. In our education setting, after a shock, schools in flagged areas follow a similar pattern of return to their equilibrium state of service delivery as their comparison schools in areas that were not flagged. Though senior managers observe the resolution of alert flags for particular administrative units, comparison to an appropriate counterfactual implies that this resolution does not seem to be due to their efforts.

It is possible that despite no overall impacts, the scheme produced significant changes in activity within the bureaucracy. We capitalize on our rich data on administrative activity to assess the impacts of flagging on key components of the public education production function. The scale of the data we have assembled allows us to estimate even small impacts with precision, painting an unusually rich picture of the impact of reforms on bureaucratic activity. We assess the financial and personnel decisions of bureaucratic managers responsible for the flagged areas. We do not observe more visits from relevant bureaucrats to affected schools, changes in their financial investments across schools, or bureaucratic transfers of teachers and head teachers. Thus overall, despite the enthusiasm for the reform of senior managers in Punjab, command and control management approaches did not motivate rank-and-file officers to change education outcomes in any substantively significant way.

We contribute to a growing literature on bureaucracy and development broadly (Finan, Olken and Pande, 2015; Besley et al., 2022), and on designing optimal incentive structures in the public sector more specifically (Banerjee et al., 2021; Ali et al., 2021). Recent (frequently experimental) papers in this literature have made the important contribution of showcasing the efficacy of various incentive schemes such as financial rewards (Muralidharan and Sundararaman, 2011; Dal Bó, Finan and Rossi, 2013; Ashraf, Bandiera and Jack, 2014; Deserranno, 2019; Leaver et al., 2021), career incentives (Khan, Khwaja and Olken, 2019; Bertrand et al., 2020; Deserranno, Leon and Kastrau, 2022), or other non-financial incentives (Ash and MacLeod, 2015; Khan, 2020; Honig, 2021). However, implementing many of these reforms at scale would require changes to the de jure environment which has been difficult to implement at scale.[4] Given the systemic nature of centralized accountability, command and control reforms are poorly suited to experimental evaluation. We present the first at-scale evidence in the economics literature on this classic pillar of Weberian bureaucracy: centralized control mechanisms.

Our findings are also relevant for the literature on the efficacy of management approaches in the public sector (Bloom and Van Reenen, 2010; Bloom et al., 2015; Rasul and Rogger, 2018; Rasul, Rogger and Williams, 2020; Banerjee et al., 2021; Ali et al., 2021; Carreri, 2021). Importantly, our study indicates that even strong centralized support for a management intervention can have passive impacts on public sector functioning. We are able to track key elements of the administrative production function precisely, supporting our null findings with evidence that the machinery of government was unmoved. Evidence on the impact of control mechanisms on public sector performance is mixed, with generally positive results for frontline settings (Olken, 2007; Hussain, 2015; Dhaliwal and Hanna, 2017; Callen et al., 2020; Duflo, Hanna and Ryan, 2012; Das et al., 2016); and less supportive evidence from experiments about administrator's motivation and performance, or those dealing with orga-

[4]By scale we mean both geographic coverage, but also temporal sustainability. Important exceptions are usually historical studies that examine major changes to civil service legislation (see for instance Xu (2018), Mehmood (2022), Aneja and Xu (2023), and Riaño (2021)). In fact, many papers examining these questions in modern bureaucracies refer to fixed de jure incentives under the Northcote-Trevelyan *system* that contain three features: competitive exam-based recruitment, rule-based promotions, and permanent civil service protected from political interference (Besley et al., 2022, p. 400). There are limited opportunities to examine how at scale changes in these impact the bureaucracy. See for instance Bertrand et al. (2020) how changes in the retirement age change career concerns in India.

nizational dynamics (Falk and Kosfeld, 2006; Dickinson and Villeval, 2008; Bandiera et al., 2021; Muralidharan and Singh, 2020). We extend this literature by providing evidence of the effects of centralized oversight on a broader administrative environment from an at-scale implementation in a large bureaucracy. By doing so, our study adds to the literature on the impacts of government-implemented schemes, which are argued to be a test of the external validity of pilot programs (Bold et al., 2018; Muralidharan and Niehaus, 2017; Vivalt, 2020) and an assessment of the most widely used public sector reforms (de Ree et al., 2017).[5]

We also add an early contribution to the nascent study of a key feature of bureaucracy: hierarchy. Though the theory of hierarchy in organizations continues to develop (Aghion and Tirole, 1997; Dessein, 2002; Chen, 2017; Chen and Suen, 2019), there are few related empirical tests in the literature. Recent evidence implies that understanding hierarchy in public organizations is critical to behavior there (Deserranno et al., 2022; Cilliers and Habyarimana, 2023). This paper shows that de-facto pressure directed through hierarchy may not engender substantial responses from public officials, however, salient senior management makes this form of incentive provision.

The paper proceeds as follows: Section 2 describes the setting of the public service we study and describes the centralized monitoring scheme. Section 3 introduces the data and presents our empirical approach. Section 4 presents the results of the evaluation of the scheme on school outcomes. Section 5 presents assessments of the scheme's impact on key elements of the education administration. Finally, Section 6 provides a discussion of our results in light of potential alternative uses of the data that was generated to run the scheme.

2 Public Education in Punjab

Punjab is Pakistan's most populous province, home to over 110 million people, half of the country's population. Twenty million are school-aged children, many attend-

[5]The paper provides a lens through which to understand the results of smaller pilots of centralized oversight, such as Callen et al. (2020), which show that flagging underperforming health facilities in Punjab positively affected health workers' attendance. However, when taken to scale, such pilots may not provide a sustainable means of managing the public administration (Banerjee, Duflo and Glennerster, 2008; Banerjee et al., 2021).

ing approximately 52,000 public schools, with 400,000 teachers (School Education Department, 2018). The scale of managing education in the province is substantial.

The province is divided into 36 districts, which are subdivided into sub-units called tehsils, further subdivided into areas of responsibility called "maraakiz."[6] There are, on average, four tehsils per district and 48 maraakiz per tehsil. Thus, on average, any district education manager has 192 administrative units to track, and each markaz official must manage an average of 20 schools.

The School Education Department is responsible for organizing and overseeing the education sector's performance. The department has two arms: district education authorities, which coordinate the implementation of public education delivery, and the Program Monitoring and Implementation Unit (PMIU), which is responsible for independently collecting and disseminating data on school performance. Both are staffed and organized separately, and monitoring is generally seen as independent of implementation.

2.1 District education authorities

Each district in the province has one district education authority which reports directly to the School Education Department. Below them, the hierarchy consists of officers for each tehsil, and assistant education officers (AEO) for each markaz. Each layer of the hierarchy is expected to manage those officers under them. AEOs are the layer of hierarchy above school principals, thus completing a multi-link chain of command from senior executive to school level.[7]

Such a layered hierarchy is not unusual in administrative settings of this scale worldwide, as the physical constraint of traveling to schools, handling administrative tasks for each, and engaging with head teachers implies a limit on the scale of any individual official's ability for oversight. By contrast, a feature of large-scale measurement in management information systems is that it can alleviate the physical constraints

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

[7]Further, schools are categorized into one of three groups: elementary education female, male, and secondary education. Our study focuses on elementary level (male and female), comprising primary schools (children aged 4 to 9) and middle schools (children aged 10 to 12). These makeup roughly 80 percent of all public schools in the province.

and centralize the ability to supervise and censure at scale. By dramatically lowering the cost of monitoring individual schools and jurisdictions, digitization of public service delivery measures has opened up the possibility of centralized management throughout the hierarchy. Such a system of monitoring the administration requires an independent administration, which we turn to next.

2.2 The PMIU

While the district education authorities are responsible for outcomes in public schools, the Program Monitoring and Implementation Unit (PMIU) is tasked with monitoring the performance of district officers. To do so, it conducts an annual census of all public schools in the province and a monthly monitoring of schools to assess key aspects of the school environment. Undertaking these duties are monitoring assistants hired to collect data.

Across the analysis period, the monitoring assistants collected performance-related data from every school on an unannounced random date every month. The assignment of monthly school inspections to monitoring assistants was randomized to limit collusion with the school staff. As we discuss in the data section, our analysis of the consistency of different data sources on schools implies that this process produced valid assessments of school performance.

Data collected by the PMIU was used for monthly and quarterly performance reports, called 'data packs.' These data packs were first generated in December 2011 and then prepared monthly. We study the period until May 2018, just before the national elections and a change in administration. The data packs reported performance at the markaz level for each district along multiple dimensions: teacher presence, student attendance, visits by district staff, and status of school facilities (electricity, drinking water, toilets, and boundary wall).[8] From September 2017, the data packs also reported scores on standardized Math, English, and Urdu tests.

The reported performance on each dimension was color-coded in the data packs based on standardized performance thresholds set by the chief minister's team. A markaz

[8] Also included the number of schools surveyed, if they were found closed, statistics by male and female schools, and recommendations about which schools to focus on to improve outcomes.

could be coded red, orange, or green, with red being the primary flag for underperformance. Figure A1 in the Appendix illustrates the color-coding. As such, an AEO (markaz-level officer) would be associated with any underperformance, although flagging was also done (in a less systematic way) at the tehsil and district level. The focus of the discussions was on markaz performance, and so that is the emphasis we follow in our empirical work, though we also provide consistent evidence for flagging at the district level.

2.3 Centralized Oversight Intervention

Using the PMIU-generated data on school performance, the chief minister of Punjab set up a centralized oversight regime for the education sector in 2011. He chaired an oversight committee and worked with the consultancy firm McKinsey International and a high-level advisor with expertise in centralized accountability.

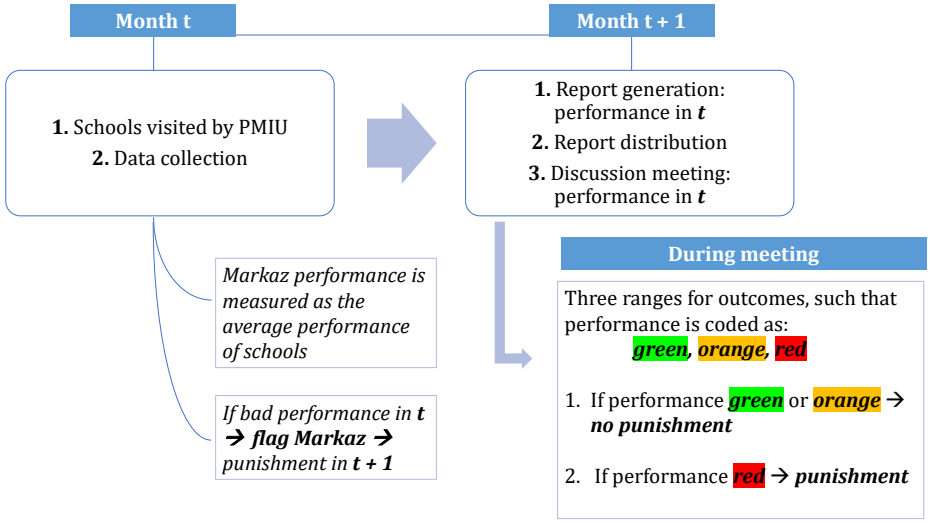
Figure 2 describes the design of the monitoring scheme. Data on all schools in the province is collected in month t. Markaz-level average performance is presented to senior managers in month $t + 1$. Maraakiz that do not reach specific (standardized) thresholds are flagged red or orange. The reports were used for senior management check-ins within the first ten days of every calendar month.

In addition, 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."[9] The senior management of the province placed substantial weight on the system, and the chief minister "had full ownership of this reform and [sent] a signal to the bureaucracy that they were to take it seriously" (Malik and Bari, 2022, p. 22).

Senior managers did not change de jure power, such as making salaries conditional on performance. Some ad hoc financial bonuses were given to district officials but not to mid-level bureaucrats. We explore whether there is evidence of staff transfers or long-term impacts on career trajectories from poor performance. We do not find any such evidence. Instead, senior management was constrained by public service rules

[9] Malik and 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."

Figure 2: Monitoring scheme structure



meant to avoid political influence. Thus, the system had to rely on de facto incentives to punish underperforming officials.

Interviews with district officials revealed that meetings mostly involved the officers flagged red getting censured in front of their peers. Quoting Malik and Bari (2022), "the red were reprimanded, and the greens were appreciated", where "The constant monitoring by the Chief Minister and the Chief Secretary played a very critical role." Officials stated that they did "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 necessarily but the point [is] that the quarterly rankings and the performance accountability caused a lot of concern."

The censoring generated incentives for district officials to motivate their subordinates, and this to those below them. The scheme intended that greater oversight by senior management would allow sanctions to serve as motivation through the chain of command. As such, the scheme relied on the interaction between measurable outcomes and personnel management. In public sector oversight models, the outputs can be reduced to observable quantities, but improvements in these still rely on multidimensional and non-contractible activities. So then, the question under evaluation is whether oversight and accountability regimes effectively motivate better personnel management throughout the hierarchy.

The political weight and international guidance ensured the scheme was effectively implemented. Reports were produced monthly from December 2011 to May 2018 as intended. To assess the data quality, we compared it with the Annual Census of Schools for the month 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 A2 in the Appendix compares both sources and shows that the overall error in reporting is low and there is a high overlap between both data sources. A comprehensive review of the data we use assesses it to be of generally high quality (World Bank, 2020).

Despite slight modifications to the scheme's structure, these elements remained at its core. As a result, the design is a demonstration case of centralized, data-informed accountability 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.

3 Evaluation methodology

3.1 Data

We used administrative data collected at the school level from December 2011 to May 2018.[10] The outcomes are monthly assessments of teacher presence, student attendance, and whether school facilities are functional. The first two are measured as the percentage of teachers/students present at the time of the visit by the monitoring assistants. The functional facilities measure the status of four types of school infrastructure: drinking water, electricity, toilets, and the boundary wall. We use an aggregate index of the share of functional facilities.

Additionally, starting in September 2017, PMIU began collecting data on student test scores in Math, English, and Urdu using standardized tests, administered by monitoring assistants to seven randomly selected 3rd-grade students in each school.

[10]The data excludes June, July, and August of each year, corresponding to summer vacations and public schools being closed.

Scores are measured as the percentage of correct answers. To understand the effect of bureaucratic behavior, we also use the data on district education staff visits to schools. We can identify each school's district, tehsil, and markaz, as well as the history of flagging across administrative tiers and units.

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

Flagging thresholds for color-coding in the datapacks were designed to be generally applicable to schools across the province, and based on the education authorities' pre-existing targets for performance measures. These targets were mostly the same across all districts and for all months of the year. In the case of student attendance, different targets were assigned across different districts and for different months of the year based on historical performance as it was felt in the case of that outcome a moving target was more appropriate.[11]

Table 1 report descriptive statistics. Panel A show that schools are relatively small, with an average of 4.6 teachers and 110 students. Roughly 3% of the schools have ever had more than 20 teachers. Those with more than 20 teachers are evenly distributed across the province. At the markaz level, Panel B shows a substantial variation in the number of schools within a markaz, broadly following differences in population size. However, the average number of schools an AEO must manage is 20, of which nearly 80% are elementary schools.

Panel A also shows descriptive statistics at the outcome-school-month level, separating between outcomes in flagged (on that outcome) and non-flagged maraakiz. Similarly, Panel B show descriptives at the outcome-markaz-month. By construction, the mean in a flagged markaz is lower than that in a non-flagged markaz. The month in which a markaz is flagged on a particular outcome, there is a drop in the mean level of that outcome. Comparison of the two sets of columns gives the order of magnitude

[11] Appendix A provides further details about the thresholds for color-coding for each indicator of interest.

of the differences. For example, flagged maraakiz have an average teacher presence of 80%, while in non-flagged maraakiz it is 93%.

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	103	2,307,637	.	.	.	.
Outcomes (%)	**No flag**				**Flag**			
Teacher presence	93	100	15	2,095,004	83	100	22	209,599
Student attendance	90	93	12	1,899,734	81	85	17	403,409
Functional facilities	93	100	16	1,875,892	84	100	22	383,125
Math score	87	92	14	824,341	67	67	21	22,212
English score	80	83	17	659,293	65	67	20	187,236
Urdu score	85	89	15	810,220	67	67	20	36,291
Panel B: Markaz-level variables								
	Mean	Median	Sd	N. Obs	Mean	Median	Sd	N. Obs
Number of schools	21	15	19	130,364	.	.	.	.
Proportion elementary	80	100	40	130,364	.	.	.	.
Outcomes (%)	**No flag**				**Flag**			
Teacher presence	93	94	4.3	95,422	80	83	7.8	8,739
Student attendance	91	92	6	89,649	80	82	7.7	14,257
Functional facilities	95	98	11	90,029	81	84	11	13,846
Math score	87	88	6.4	60,069	65	66	4.9	2,100
English score	80	80	6.3	49,451	64	66	5.1	12,718
Urdu score	85	86	6.4	59,375	65	67	5.1	2,794
Panel C: District level variables								
	Top 5				**Bottom 5**			
Outcomes (%)	Mean	Median	Sd	N. Obs	Mean	Median	Sd	N. Obs
Overall score	94	95	3.8	70	78	78	10	70
New position	.077	0	.27	504	.083	0	.28	504

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

In addition to the monthly flagging of AEOs/maraakiz, the districts were ranked each quarter. The ranking was based on an overall score of the performance in the

previous months.[12] Panel C in Table 1 shows descriptive statistics for districts in the top/bottom positions. Bottom districts report a lower mean in the score. Panel C also shows the percentage of districts that entered the top/bottom five positions in each period. There is a relatively small number of cases where new districts fell into the top (7.7%) or bottom (8.3%) positions, suggesting a high degree of persistence in the ranking status.

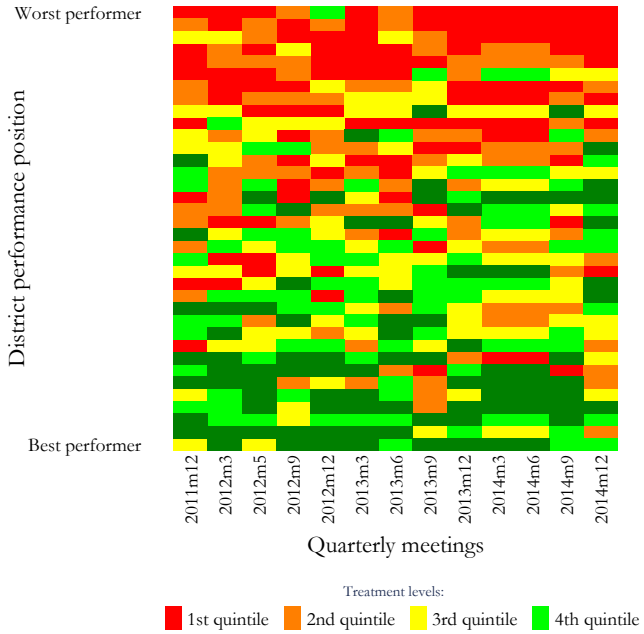
Figure 3 presents this persistence graphically. For each quarterly meeting, we color-coded the quintile in which the district fell in the overall score distribution. The districts in the higher quintiles tend to maintain their high position in the ranking. In contrast, the districts in the lowest quintiles remained in last. The figure thus presents a descriptive sense that the flagging did not motivate poor performers sufficiently for their overall rankings to change.

A feature of the intervention environment is that almost all maraakiz were flagged at some point, and yet some districts and maraakiz remain systematically at the bottom of the distribution. Evidence from other settings indicates that education (and other environments) face structural constraints to improving outcomes (World Bank Group, 2018). However, they are also exposed to shocks (such as teachers getting sick) that substantially shift 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 performance levels.

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

[12] 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.

Figure 3: Distribution of quintiles of district performance



Note: This figure illustrates for each quarter the quintile of the overall district score distribution in which each district fell. District scores are measured based on the aggregate performance of teacher presence, student attendance, and functional facilities in each quarter. The figure ranks the districts based on their average performance of all the periods, such that the worst performing district at all times appears first.

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	.96	.67	.72	1.5	7.4	51,534
Student attendance	13	1.3	.77	.69	1.5	9.6	51,507
Functional facilities	17	4.5	1.7	.46	.64	16	50,501
Math score	5.4	1.1	.81	.79	1.8	6.3	37,537
English score	5.8	1.4	1.1	1.2	3	8.2	37,536
Urdu score	5.5	1.3	.93	.92	2	7	37,536

Notes: The unit of observation for outcomes is presented at the school level. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. Each quintile is calculated separately based on the mean level of performance for each variable. The table shows the standard deviation for each school-level variable quintile.

3.2 Empirical strategy

To estimate the effect of the centralized accountability system on educational outcomes, we followed Cengiz et al. (2019) and Baker, Larcker and Wang (2022) to build a stacked dataset to avoid biases driven by the time-varying nature of the treatment (De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021). The stacking consists of creating event-specific datasets for identifying control units that have not been treated during a specific period. The process is described in Figure A3 in the Appendix. The result is a dataset 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.[13]

3.2.1 Markaz flagging

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

$$\begin{aligned} Y_{smdte} = \gamma_1(T_{mde} \times Flag_{te}) + \gamma_2(T_{mde} \times Punish_{te}) + \\ \beta(T_{mde} \times AfterFlag_{te}) + \alpha_{mde} + \lambda_{te} + dt + \epsilon_{smdte} \end{aligned} \tag{1}$$

Subscripts s, m, d, t are for school, markaz, district, and time. All of the components are indexed at the event panel e. Y_{smdte} is the outcome for school s, within markaz m, in district d. T_{mde} equals 1 for schools in a flagged markaz m. $Flag_{te}$ equals 1 for the period data is collected and the flag is defined. $Punish_{te}$ equals 1 after the

[13] The core empirical exercise we conduct in this paper uses specific features of the flagging system to estimate rigorous identification of its effects. However, in Appendix B.1 we also assess whether the introduction of the scheme itself created large changes in the trends of public education outcomes. We do so in three ways. First, assessing whether outcomes trended similarly before and after the introduction of the scheme in Punjab versus other territories in Pakistan. Second, whether the first flagging of any jurisdiction had a particular impact on its trajectory. Third, whether the first flagging of a jurisdiction in a district had any impact on the wider trajectory of schools there. We find no evidence on any of these margins: the introduction of the scheme did not affect the trajectory of school outcomes. This alleviates the concern that the relevant responses of bureaucrats to the scheme happened before (in expectation) or on impact. Such a coordinated and widespread response seems intuitively unlikely in a large and disparate environment.

flagging, where the oversight committee meets and the accountability intervention occurs. $AfterFlag_{te}$ equals 1 after the punishment phase where we assess the intervention impact. α_{me} is for markaz fixed effects to control for constant characteristics of maraakiz, and λ_{te} is for time fixed effects to capture time-specific shocks. We include dt –a district binary and linear calendar index– to absorb district linear time trends. ϵ_{smdte} is the error term clustered at the markaz level (treatment level). In our main specifications, we stack for four pre-periods and seven post-periods.

Figure 4 presents the evolution of our outcomes in relative time, anchored on periods of flagging. Solid lines are schools in flagged maraakiz. Dotted lines are schools in non-flagged maraakiz. We present two dynamics: one that uses all schools (the blue lines) and one that uses only those that are "close" to the threshold for flagging (red lines). We highlight three periods corresponding to the month in which the data is collected and the flag is defined, 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.

We observe that treated and control units follow similar paths just before the flagging. In the month of flagging, the average school in a markaz that gets flagged suffers from a shock, contributing to the markaz being selected for treatment.[14] Thus, the treated units would not have followed the same transition as control units without the treatment, and the conditions for causality would be violated. To address the parallel trends violation, we follow Rambachan and Roth (2022) and redefine the base period as the one just before the negative transitory shock occurs (relative time -1).

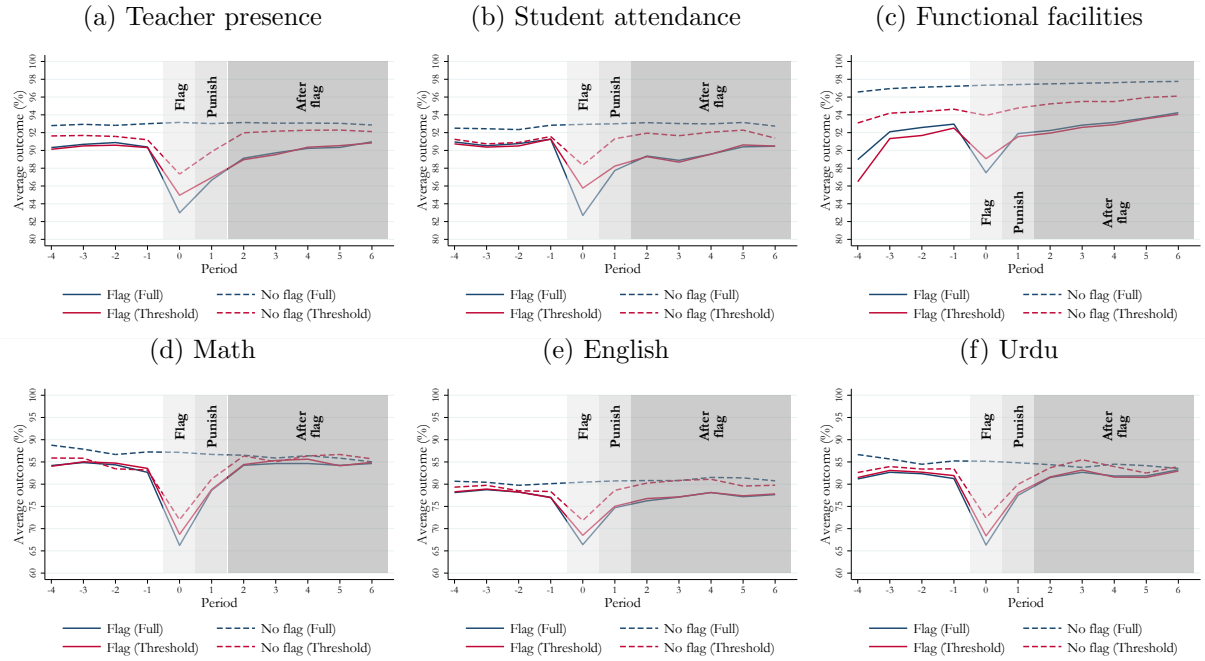
To further account for the negative shock, we build a sample of comparable schools around the flagging thresholds for each outcome (plotted in red). We follow Calonico, Cattaneo and Farrell (2020) to identify the maraakiz within an optimal bandwidth on either side of the flagging threshold in time 0. We obtain optimal bandwidths separately for each event panel to build an stacked-threshold sample.[15] As can be observed, regardless of the sample, the transition of outcomes typically reverts to the

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

[15]The threshold sample consists of 16% of observations of the full sample for teacher presence and student attendance, 9% for functional facilities and Urdu scores, 5% for Math scores, and 23% for English scores.

pre-shock levels.

Figure 4: Evolution of school outcomes in relative time - markaz flagging



Note: The figure presents the average evolution of schools in flagged (continuous line) and non-flagged (dashed line) maraakiz. Flagging is based on the outcome variable in focus. Blue lines represent the full sample. Red accounts for the threshold sample that is "close" to the flagging threshold. Relative time is divided into: *Flag*: period where information is collected and maraakiz are flagged; *Punish*: period where the reports are distributed and oversight meetings are held; *After flag*: periods after the meeting.

Then, γ_1 absorbs the effect of the negative transitory shock, and γ_2 captures the immediate recovery in the punishment period. β would estimate the effect of flagging on school performance after the shock. If flagging leads to higher 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.

3.2.2 District ranking

One concern is that markaz flagging might be less salient when the rest of the district performs well. We complement our core strategy with analysis at the district level.

Above we noted that in quarterly oversight meetings, districts were ranked according to the aggregate performance in the prior quarter. Though we are far less powered 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, and additionally look at the interaction between district and markaz flagging.

We stack for four pre-periods and three post-periods as district meetings happen quarterly. We use as event time each month in which a meeting happened. Flagged units are defined as the schools in districts that were at the bottom/top of the ranking during the meeting in period 0. District rankings do not systematically receive a negative shock before the meeting, and thus do not require corrections for related self-selection and reversion to the mean. However, for consistency, we define -1 as the base period and build a threshold sample of the five districts closest to the treated five at the top/bottom to represent a threshold comparison. We estimate the effect of district ranking with the equation below:

$$\begin{aligned} Y_{smdte} = \gamma(Position_{de} \times Meeting_{te}) + \\ \beta(Position_{de} \times AfterMeeting_{te}) + \alpha_{de} + \lambda_{te} + \epsilon_{smdte} \end{aligned} \tag{2}$$

$Position_{de}$ equals 1 for schools in bottom/top districts d. $Meeting_{te}$ equals 1 for the period when the quarterly meeting happens, so γ absorbs any immediate effect of the meeting. $AfterMeeting_{te}$ equals 1 for the months after the meeting, so β estimate the persistent effects of the flagging. α_{de} are district fixed effects and λ_{te} are time fixed effects. ϵ_{smdte} is the error term clustered at the district level. Interactions between this specification and the above markaz-level specification are natural extensions to these equations.

4 Results

4.1 Markaz flagging

Figure 5 reports the event studies for each outcome variable we study. The y-axis reports β coefficients in percentage point differences. The blue line is the full sample, while the red is the threshold sample. The event studies show that the pre-trends are not significant and are small in magnitude. Thus, the parallel trends assumption is plausible. As can be seen, most of the coefficients in both samples are statistically equivalent to zero at the 95% level in the *After flag* period, indicating null impact of the flagging. The full sample estimations exhibit a larger relative negative shock measured in period 0, but even this is almost recovered by the first *After flag* period.

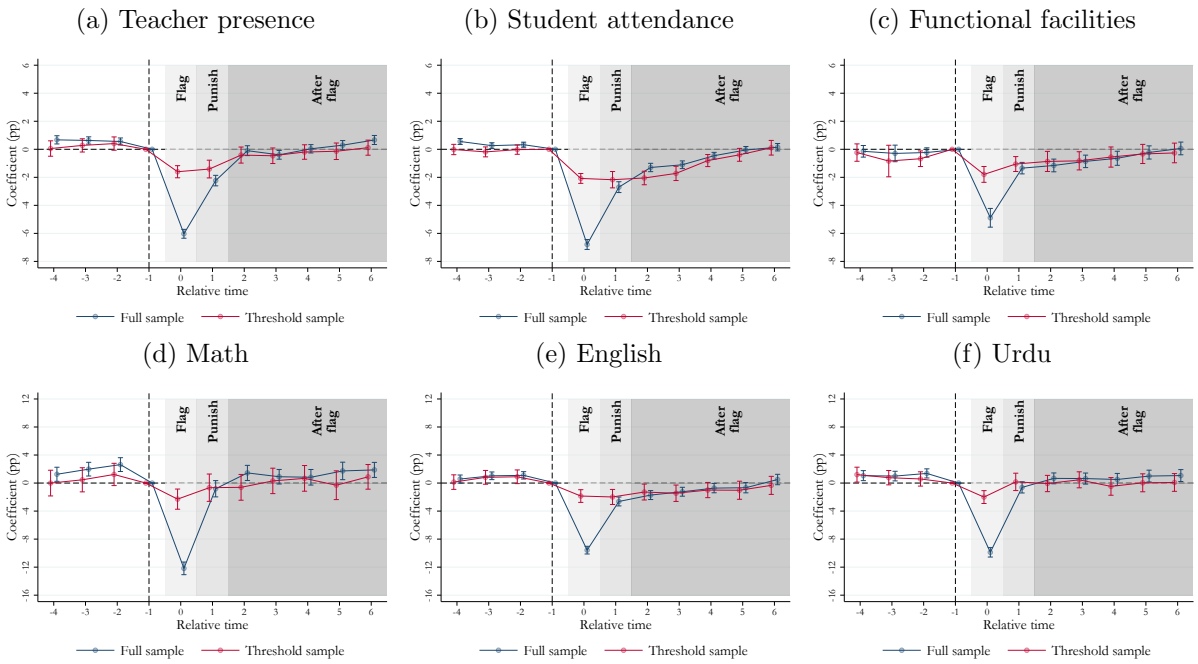
We in fact see flagged schools taking longer than their equivalent non-flagged schools to return to their pre-existing levels. In particular, the coefficients related to student attendance (panel b) and English scores (panel e) take longer to reach the pre-shock level in flagged schools, though the magnitude of the effects are small. This is likely due to the fact that treatment schools have a marginally stronger shock in the outcome variable, and they may naturally have a more extended transition back to equilibrium.[16]

Table 3 presents the results of estimating equation 1. The first column for each variable reports the full sample, and the second shows the threshold sample. Panel A reports outcomes relating to school functioning. They are always negative and significant coefficients in the *Flag* and *Punish* periods for flagged 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 a reversion to the mean effect.

The coefficients for the *After flag* period (corresponding to β) are significant in both samples for teacher presence and student attendance. The coefficients are small (almost zero) compared to the mean of the dependent variable, but negative rather than

[16] As a robustness check, Figure B5 reports the event studies for a stacked dataset with fewer post periods to test the sensibility of the results to an arbitrary number of periods. Results follow the same trends in both cases.

Figure 5: Event study - flagging effect on performance



Note: This figure presents results from estimating event studies based on equation 1 using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

positive. As observed graphically, the negative effect can be interpreted as a persistence of the negative shock. Panel B of Table 3 presents the results for the student test score variables. We note that the sample size is smaller here, given the reduced time frame for which we have these measures. We observe the same pattern of results as in Panel A. The results imply that the oversight scheme had no impact on school functioning nor student outcomes, but rather that flagged and non-flagged schools facing a similar shock returned to equilibria at roughly the same rate, and certainly did not improve disproportionately beyond their pre-existing levels.

To assess the robustness of these results, we present a series of additional specifications in Appendix B. We plot the estimate for each coefficient from a stacked data set including t additional periods to further test the stacked structure (Figure B4). We

Table 3: Monitoring effect on performance - markaz flagging

Panel A: School outcomes						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-6.51***	-1.81***	-7.13***	-2.11***	-4.73***	-1.39***
	(0.15)	(0.19)	(0.18)	(0.16)	(0.34)	(0.28)
T×Punish	-2.71***	-1.63***	-3.06***	-2.24***	-1.19***	-0.66**
	(0.18)	(0.27)	(0.19)	(0.29)	(0.18)	(0.29)
T×After flag	-0.40***	-0.47***	-0.98***	-1.18***	-0.43***	-0.23
	(0.098)	(0.16)	(0.081)	(0.15)	(0.16)	(0.29)
N. of obs.	6,979,566	490,950	4,964,842	562,661	7,314,616	392,052
Mean Dep. Var. before	92.9	87.3	91.8	87.2	97.4	93.9
R^2	0.032	0.036	0.10	0.098	0.070	0.069
Panel B: Student scores						
Dependent variable:	Math		English		Urdu	
T×Flag	-13.7***	-2.74***	-10.3***	-2.34***	-10.7***	-2.63***
	(0.35)	(0.56)	(0.22)	(0.30)	(0.26)	(0.34)
T×Punish	-2.31***	-1.11	-3.34***	-2.47***	-1.48***	-0.46
	(0.46)	(0.82)	(0.28)	(0.46)	(0.36)	(0.55)
T×After flag	-0.14	-0.28	-1.53***	-1.55***	-0.087	-0.61*
	(0.28)	(0.52)	(0.20)	(0.30)	(0.22)	(0.35)
N. of obs.	2,182,972	53,066	804,855	146,692	1,936,332	119,016
Mean Dep. Var. before	86.9	71.7	78.1	70.4	84.5	71.7
R^2	0.100	0.15	0.065	0.069	0.10	0.13
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

Notes: Results from estimating equation 1. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. The flagging and threshold sample are based on the studied outcome. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. 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. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

assess the effects of using fixed effects that absorb the history of markaz flagging (Table B1) and of changing our assumptions on the persistence in the impact of flagging (Table B2). We present alternative difference-in-differences estimators (Figure B6

and B7). We estimate the results for the 'orange' flagging threshold (Figure B8) and for flagging at the tehsil level, the layer of hierarchy above the markaz (Figure B9). We assess the impact of centralized accountability separately for each month in which the scheme was implemented (Figure B10). We also investigate the possibility of public officials anticipating the flagging (Figure B11). In all cases, our results are qualitatively the same.

One possibility is that the system was not intended to improve student outcomes but rather to serve political ends. We therefore assess whether flagging had differential impacts across political environments. In Appendix B.5 we identify political alignment following Callen, Gulzar and Rezaee (2020) and use a difference-in-differences strategy to assess the effect of being in a politically aligned markaz. We compare aligned/non-aligned markaz, before/after the 2013 elections in places with high political competition (close elections). While we find no effect of political alignment on the probability of flagging, there are small effects of alignment on student attendance itself but no consistent effects elsewhere.

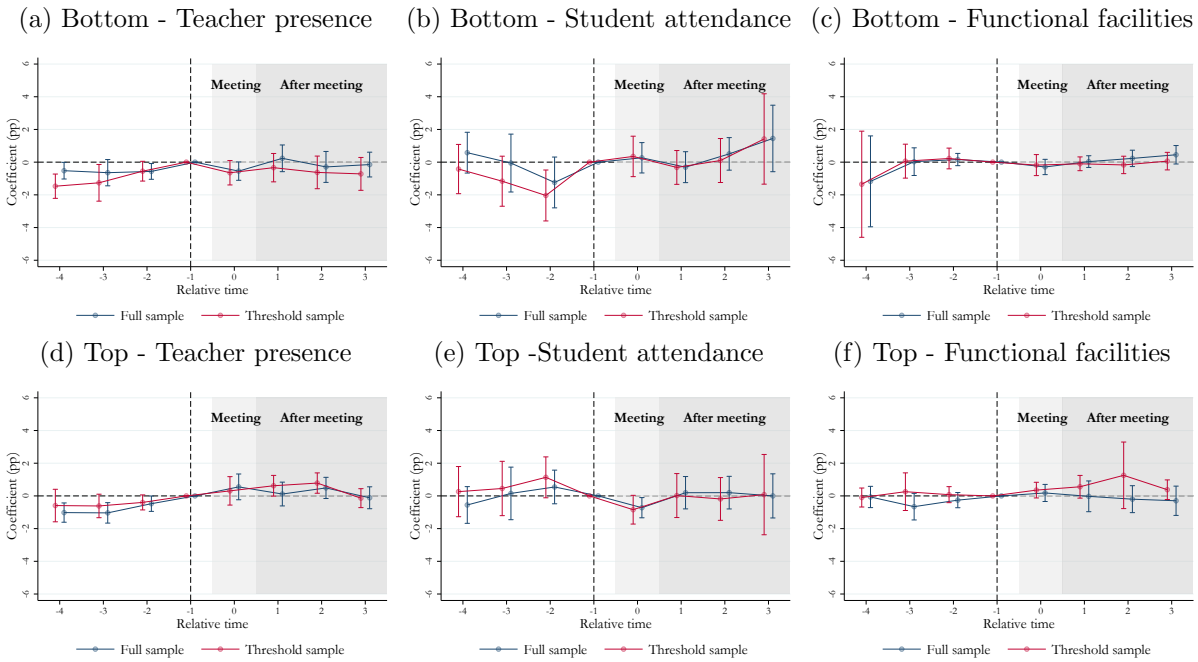
4.2 District ranking

To complement our main analysis, we assess the impact of being the top/bottom performing districts at quarterly oversight meetings. We restrict our analysis to measures of school functioning. Figure 6 presents the event studies for top/bottom performing districts. The figures illustrate that no pre-periods appear significant, suggesting the plausibility of the parallel trends assumption. The *After flag* period indicates that flagging once again has no significant effects on school functioning or outcomes.

Table 4 reports the treatment effects. Panel A for schools in the bottom districts shows that we detect a small but significant increase of 1.3 percentage points in student attendance for the threshold sample. Panel B for schools in top districts shows that being in it leads to a slight increase in teacher presence after the quarterly meeting. However, the coefficients are small in magnitude relative to the mean of the dependent variable before the meeting (91% in the full sample and 91.9% in the threshold sample). Hence, there is no evidence of significant increases in performance

due to centralized monitoring of higher-level managers from the district-level rankings. The findings are consistent with the descriptive statistics in Panel C of Table 1 and Figure 3, 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.

Figure 6: Event study - district ranking effect on performance



Note: This figure presents the results from estimating an event-study based on equation 2, using -1 as base period, comparing schools in top/bottom districts against schools out of the top/bottom districts. Bottom is for the schools in bottom five districts in the quarterly meeting. Top is for the schools in the Top five districts in the quarterly meeting. Blue line accounts for the result on the full sample, while the red accounts for the results using the threshold sample, including the schools in the five districts closer to the five in the bottom/top. *Meeting* is for the period of the quarterly meeting. Error bars at the 95 percent level are presented for each coefficient.

Despite finding zero overall impacts of flagging at the district level, we tested the impact of the interaction between district-level and markaz-level flagging. We hypothesize 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 2, 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 quarterly meeting. Appendix Table B3 reports the results of the heterogeneity analysis. Panel A reports the

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
	(0.33)	(0.36)	(0.56)	(0.61)	(0.52)	(0.53)
N. of obs.	3,063,835	583,417	3,063,410	583,248	3,009,844	565,920
Mean 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.	3,111,642	682,461	3,111,048	682,369	3,036,557	672,780
Mean 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: Results from estimating equation 2. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including the schools in the five districts closer to the five in the bottom/top. The bottom/top status and threshold sample are based on the aggregate district performance. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. *Bottom* equals 1 for schools in the bottom five districts and Top equals 1 for the schools in the top five districts on the date of the quarterly meeting. *Meeting* equals 1 in the period of the quarterly meeting. *Mean. Dep. Var before* shows the average outcome in the non-top/bottom districts before the meeting occurs. Standard errors clustered by district, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

results for the bottom districts, while Panel B reports the results for the top districts. The triple interactions for none of the panels, variables, and samples show positive and significant results, suggesting no major interaction between flagging district- and markaz-level performance.

4.3 Does punishment change the trend of recovery?

The recovery to pre-treatment means is a combination of mean reversion and the impact of the punishment period. A key advantage of the frequency of our data is that we can separately examine the impact of punishment beyond the regression to the mean trends in the outcomes. To do so, Figure 7 plots over time impacts on first-differenced outcomes that are reported above in Figure 5.

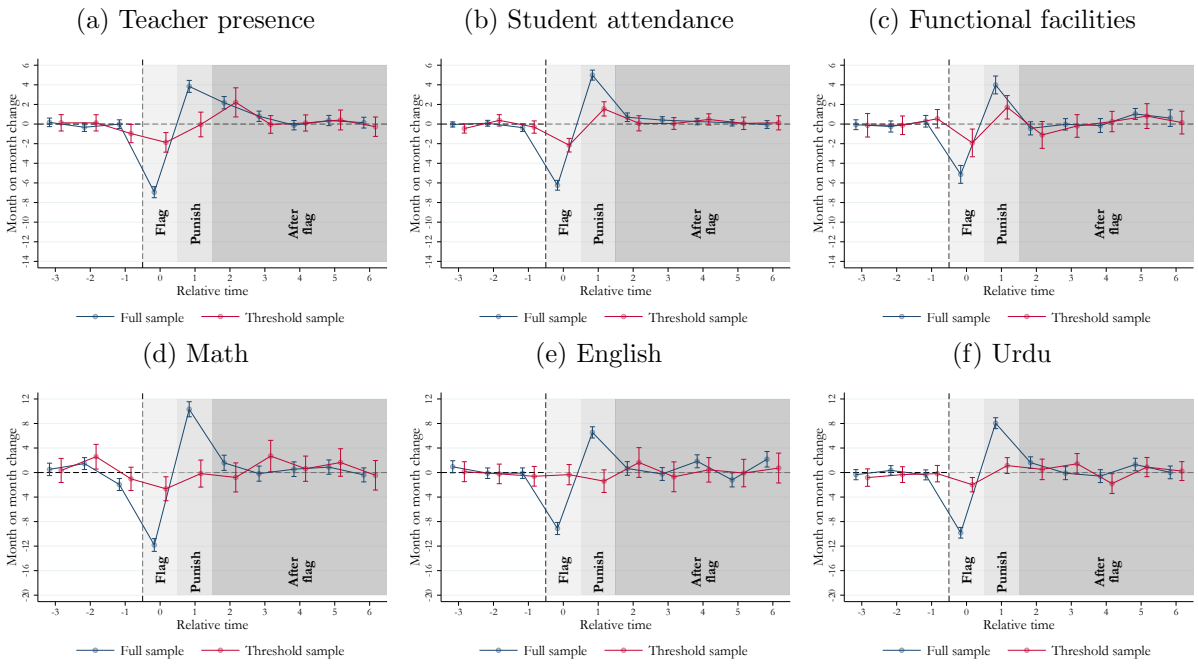
We can see that there exists a negative shock during the flagging month ($t = 0$). This negative shock is followed by a quick recovery in the month where punishment occurs ($t = 1$). If it were the case that punishment, where top down accountability occurs, was contributing to an improvement *beyond* the pre-existing path of recovery, we would expect the coefficient in period $t = 2$ to be larger than the coefficient in period $t = 1$ as the path to recovery would have accelerated.

We find evidence for the efficacy of punishment only in the case of teacher presence (panel a), where there is a small precisely estimated effect on the first differenced outcome (p-value of 0.06). This shows that the rate at which teachers return to schools is increased in the first month after flagging by 2 percentage points. From month 2 onwards, we see no difference between flagged and non-flagged schools. The results for flagging on other outcomes are all indistinguishable from zero, suggesting that punishment is not bringing any further improvement in the rate of recovery. Taken together, these results show that there is an impact of top-down accountability but it is small in magnitude and only occurs on the immediate next step on the causal chain.

5 Impacts on the machinery of government

Despite finding no impacts of the centralized accountability scheme on schooling outcomes, we can use the data we have collected to investigate if there were effects on other bureaucratic activities that we would expect to observe if the bureaucracy had been motivated to respond to the flagging. Specifically, we can analyze administrative action in terms of both personnel and financial resources, the two key inputs to effective government functioning. We look at bureaucratic effort through monitoring

Figure 7: Punishment Period vs Reversion to Mean - Month on Month Changes



Note: This figure presents results from estimating month-by-month coefficients based on equation 1 on the sample of maraakiz that have not fully recovered from the negative shock in the punishment period. The specification compares schools in flagged and non-flagged maraakiz in consecutive months. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient. We report the p-values for a one-sided test for the coefficient of relative time 2 (after flag) being greater than the coefficient of relative time 1 (punishment) in the threshold sample: Panel (a) –0.06–. Panel (b) –0.99–. Panel (c) –0.99–. Panel (d) –0.62–. Panel (e) –0.08–. Panel (f) –0.6–.

visits to affected schools, the movement of staff, and impacts on promotions. We also look at changes in school budgetary resources and the nature of expenditures at the school level.

Oversight visits. A natural immediate response by public officials flagged for poor performance would be to visit poorly performing schools to undertake diagnostic and remedial work on whatever area of school functioning had been flagged. School visits are a standard part of the AEOs work program and a mechanism to resolve issues that schools face in functioning effectively. We explore whether the flagging led to an increase in visits to (affected) schools. Table B5 in the Appendix reports the results

of each flagging on this measure of bureaucratic effort. The 'visited schools' measure equals 1 if the school received a visit from the relevant AEO. The coefficients of *Flag* and *Punish* periods account for changes in the probability of receiving a visit, given the negative shock. The flagging has no significant effects on bureaucratic visits to schools. The coefficients for the *After Flag* period are never significantly positive for both samples in none of the variables. Table B6 shows the results in samples with specific characteristics to explore if bureaucrats gamed the system by strategically visiting bigger, worst performing, or most missing teachers schools. The results are small or non-significant.

School Budget Utilization. Another response by public officials is to channel budgetary resources to support struggling schools. We explore the relationship between flagging and the schools' resources by aggregating the panel at the year level and counting the number of times each school was in a flagged markaz. We used a panel regression with markaz and year fixed effects, and district-time trends to obtain estimates of the impact of the number of times flagged in a year on the amount of funds received and the expenditures undertaken by the schools in the next year. Further, to address potential endogeneity from resources assignment also affecting the flagging status, we use an instrumental variables approach and exploit the random position of the markaz around the arbitrary threshold. We instrument the number of times flagged by whether the schools were in a markaz flagged when staying within the threshold sample.

Panel A of Table B7 in the appendix shows the results for each flag type on the amount of funding given by the government and the reported expenses at the school level for a year. For teacher presence, one more flag in the previous year is associated with an increase of 6% in non-government funds (those received from non-government sources such as parents), and one more flag on functional facilities increases non-government funds by 3%. For student attendance, one more flag leads to a 7% increase in government funds received by the school and a 3% rise in expenditures. The rest of the coefficients are small in magnitude and broadly insignificant. Panel B reports the results from an IV estimation strategy where we instrument the number of times flagged by the distance to the flagging threshold, which is akin to fuzzy RD setup. In this setup, we find no evidence of a response to any of the flagging in either the funds received by a school or its expenditures. Overall, there seems little

systematic evidence that the flagging shifts budgetary resources or expenditures.

Transfers and Postings. Public officials can also intervene in the management of schools through the labor market by moving head teachers, or district officials across schools or districts in response to flagging. We study the rotation of officials at the school and district level, measuring rotation as a variable that equals 1 if the public official reported in period t is different from the one reported in $t-1$. First, we explored whether the markaz flagging induced a higher rotation of head teachers, as AEOs might use it to improve school performance within their administrative unit. We used equation 1 with rotation of head teachers as a dependent variable. We thus estimated the effect of being flagged on the probability of observing head teacher rotation. Overall, Appendix Table B8 shows no significant changes in the probability of rotation of head teachers, except from math and english scores, for which we found a lower probability of rotation in the after-flag period.

Second, we used equation 2 at the district level to observe the rotation in district managers themselves. Because the district officer is a district attribute, we aggregated the data at the district level. Panel A of Appendix Table B9 reports the results for bottom-performing district, and Panel B for 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 a higher rotation of officers.

Finally, we explored for the district-level officers whether being in charge of a top/bottom district was related to whether they held a higher/lower-ranked position at the end of the scheme. In other words, whether the success of the districts in which they were in charge had any impact on their long-term progress through public service. We obtained data on the current employment of public officers in charge of a district between 2011 and 2015 and generated a ranking of the importance and status of each role. Appendix A.5 details how we constructed the rankings of district officer positions. We also calculated the months they were in charge of a top/bottom district. Then, we estimated a simple regression correlating the ranking of the current employment and the number of months they were in charge.[17] No coefficient is significant. However, we do have a relatively small number of observations and observe

[17]Regressions use bootstrapped standard errors to account for the low number of observations.

that the bottom (top) districts are negatively (positively) correlated with the rank of the current position of the public official.

Overall, there is no consistent evidence that the central accountability scheme induced any substantive impacts on how the government functioned in bureaucratic effort, budget, or public sector labor market, a result consistent with the null impacts that the scheme had on the targeted variables.

6 Discussion

Centralized command of the public administration, typically with few related changes in the de jure incentive structure, has been a dominant approach to the management of the public sector (Finer, 1997; Education Commission, 2023). The rise of public service digital information systems has brought greater attention to the efficacy of this approach. As centralized analytical units have fed substantial volumes of data to senior managers, governments have been keen to showcase their responsiveness to this data through top-down methods of controlling service delivery. Despite the prevalence of this approach to managing government throughout history, as well as its continued implementation at scale worldwide, there have been limited evaluations to date on its efficacy.

We analyze the effectiveness of 'command and control' in government administration by evaluating a system from Punjab province in Pakistan that alerted senior government managers to poorly performing school districts. Despite flagging of poor performance leading to de facto accountability along the bureaucratic hierarchy, the scheme had no substantive impacts on schooling outcomes across any targeted outcome. By assessing the activities of public officials throughout the chain of service delivery, we find that this system had no impact on any aspect of government functioning beyond a slightly faster return of teachers to schools flagged as having low teacher attendance. Our data allow us to make these claims with a high degree of precision. Taken together, our results suggest that centralized command and control management approaches struggle to effectively manage unpredictable delivery environments. Such findings are consistent with emerging literature on large-scale incentive provision in the public service (see introduction).

An obvious caveat to our findings is that de jure incentives were not changed, and thus it could be argued that we would not expect to see responses by rational economic actors. However, widespread literature on the personnel economics of the state has documented the challenges to sustained changes in formal public sector contracts (Banerjee et al., 2021) and the dominance of de facto public sector incentive schemes implemented in reality (Schuster et al., 2023). As such, a frontier of that literature is to understand how de facto incentives (such as top-down accountability) may or may not improve service delivery outcomes.

Our identification strategy focuses on the impacts of the flagging of underperforming schools. There may have been larger benefits of the scheme, such as an immediate accountability effect or wider learning across the system upon its introduction. However, assessing the immediate impacts of the scheme using a range of approaches, we also fail to find evidence that its introduction substantially shifted outcomes. We also do not see any broad shift in the ranking of districts across the province, such that any learning did not improve the performance of the weakest performers. Rather, the relative rankings of school performance persisted. More broadly, a threshold-based approach to performance measurement is unlikely to be the most relevant method for a system to maximize learning, given its narrow lens. Alternative reporting based on the same data may have captured relative progress better.

What do these findings imply for large-scale data collection in the public sector? However detailed data-collection, management information systems struggle to document the full extent of many modern public service environments. As such, there have long been calls for autonomy for effective frontline service managers in related literatures (Simon, 1983; Dixit, 2002). At the same time, large-scale datasets combined with modern analytics have been shown to be a powerful means for estimating important structural elements of the public sector production function (Fenizia, 2022; Best, Hjort and Szakonyi, 2017). This would suggest that there is utility from taking an approach that builds on the comparative advantage of large-scale data analysis in estimating more permanent parameters of the education production function rather than variables that are potentially vulnerable to short-run stochastic shocks.

As an illustration of the power of large-scale data in the case of Punjab, we use the PMIU data to estimate the impacts of head teacher quality on the same outcomes that

the centralized accountability system focused on. We follow Fenizia (2022) in using an AKM-model (Abowd, Kramarz and Margolis, 1999) of head-teacher productivity (Card, Heining and Kline, 2013). We identify three important insights. First, head teachers have different levels of added-value across distinct areas of school functioning, with some better at inducing teacher presence, and others better at improving test scores. Second, the rotation of head teachers across schools can have substantial impacts on school outcomes. Overall, a one standard deviation increase in head teacher quality accounts for approximately 3% improvement in the corresponding outcome of the average school in our sample. Third, by using this information to optimally allocate head teachers to schools that are most in need of a particular set of skills, we find that PMIU could have raised levels of teacher presence by 19 percentage points in schools that were performing below the median on that margin.

Combining this illustrative analysis with our results indicates that centralized management of service-delivery through high-frequency monitoring and related control methods is an inefficient use of information management systems in the public sector. We find no evidence that this statement is mediated by features of the targeted outcome, with the 'command and control' scheme we study having no impacts along any point on the causal chain: from monitoring and budgetary allocation, facility construction and maintenance, to student and teacher presence at schools. However, insights using the data resulting from high-frequency monitoring can be powerfully used to identify structural parameters of the education production function that no official within the public administration could generate independently.

Moreover, complementing large-scale and high-frequency data collection with appropriate counterfactual analytics ensures that limited public resources are spent judiciously. We estimate, using only that data which PMIU would have had access to during the rollout of the scheme, that the limited impacts of the system could have been detected within months of it starting. Figure B12 in the Appendix plots the after-flag β coefficients from equation 1 using only the data available up to month t.[18] As such, we mimic the analysis that the government could have undertaken during the scheme's operation.[19] The results are a long string of null or negative coefficients

[18] In the first month, we use data from the first month only. In the second, we use data from the first two months, and so on.

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

that would have been quickly perceptible to an analyst. Financial and personnel resources, and the attention paid to the scheme, could have been repurposed to other, potentially more effective, policies.

In conclusion, our paper provides a detailed evaluation of the concerns with centralized accountability systems debated in the literature (Kane and Staiger, 2002; Besley and Coate, 2003; Bardhan, 2002; Bó et al., 2021). Our results support the perspective that oversight and control approaches fail to induce changes throughout a public sector hierarchy. However, re-purposing the data that underlies an oversight scheme for analytical purposes related to structural determinants of public sector effectiveness has much greater promise (Lang, 2010; Staiger and Rockoff, 2010).

References

Abowd, John M, Francis Kramarz, and David N Margolis. 1999. "High wage workers and high wage firms." *Econometrica*, 67(2): 251–333.

Aghion, Philippe, and Jean Tirole. 1997. "Formal and Real Authority in Organizations." *Journal of Political Economy*, 105(1): 1–29.

Ali, Aisha J, Javier Fuenzalida, Margarita Gómez, and Martin J Williams. 2021. "Four lenses on people management in the public sector: an evidence review and synthesis." *Oxford Review of Economic Policy*, 37(2): 335–366.

Aneja, Abhay, and Guo Xu. 2023. "Strengthening State Capacity: Civil Service Reform and Public Sector Performance during the Gilded Age."

Ash, Elliott, and W. Bentley MacLeod. 2015. "Intrinsic Motivation in Public Service: Theory and Evidence from State Supreme Courts." *The Journal of Law and Economics*, 58(4): 863–913.

Ashenfelter, Orley. 1978. "Estimating the effect of training programs on earnings." *The Review of Economics and Statistics*, 47–57.

Ashenfelter, Orley C, and David Card. 1984. "Using the longitudinal structure of earnings to estimate the effect of training programs."

Ashraf, Nava, Oriana Bandiera, and B Kelsey Jack. 2014. "No margin, no mission? A field experiment on incentives for public service delivery." *Journal of public economics*, 120: 1–17.

Baker, Andrew C, David F Larcker, and Charles CY Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics*, 144(2): 370–395.

Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat. 2021. "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats*." *The Quarterly Journal of Economics*, 136(4): 2195–2242.

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

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

Barber, Michael. 2013. "The Good News from Pakistan." Reform, London.

Bardhan, Pranab. 2002. "Decentralization of Governance and Development." *Journal of Economic Perspectives*, 16(4): 185–205.

Bertrand, Marianne, Robin Burgess, Arunish Chawla, and Guo Xu. 2020. "The glittering prizes: Career incentives and bureaucrat performance." *The Review of Economic Studies*, 87(2): 626–655.

Besley, Timothy, and Stephen Coate. 2003. "Centralized versus decentralized provision of local public goods: a political economy approach." *Journal of Public Economics*, 87(12): 2611–2637.

Besley, Timothy, Robin Burgess, Adnan Khan, and Guo Xu. 2022. "Bureaucracy and Development." *Annual Review of Economics*, 14(1): 397–424.

Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2017. "Individuals and organizations as sources of state effectiveness." National Bureau of Economic Research.

Bloom, Nicholas, and John Van Reenen. 2010. "Why Do Management Practices Differ across Firms and Countries?" *Journal of Economic Perspectives*, 24(1): 203–24.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen. 2015. "Does Management Matter in schools?" *The Economic Journal*, 125(584): 647–674.

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

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2018. "Experimental evidence on scaling up education reforms in Kenya." *Journal of Public Economics*, 168: 1–20.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.

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

Callen, Michael, Saad Gulzar, and Arman Rezaee. 2020. "Can political alignment be costly?" *The Journal of Politics*, 82(2): 612–626.

Calonico, Sebastian, Matias D Cattaneo, and Max H Farrell. 2020. "Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs." *The Econometrics Journal*, 23(2): 192–210.

Card, David, Jörg Heining, and Patrick Kline. 2013. "Workplace heterogeneity and the rise of West German wage inequality." *The Quarterly journal of economics*, 128(3): 967–1015.

Carreri, Maria. 2021. "Can good politicians compensate for bad institutions? Evidence from an original survey of Italian mayors." *The Journal of Politics*, 83(4): 1229–1245.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The effect of minimum wages on low-wage jobs." *The Quarterly Journal of Economics*, 134(3): 1405–1454.

Chaudhry, Rastee, and Abdullah Waqar Tajwar. 2021. "The Punjab Schools Reform Roadmap: A Medium-Term Evaluation." *Implementing Deeper Learning and 21st Century Education Reforms: Building an Education Renaissance After a Global Pandemic*, , ed. Fernando M. Reimers, 109–128. Cham:Springer International Publishing.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *American Economic Review*, 95(4): 1237–1258.

Chen, Cheng. 2017. "Management Quality and Firm Hierarchy in Industry Equilibrium." *American Economic Journal: Microeconomics*, 9(4): 203–44.

Chen, Cheng, and Wing Suen. 2019. "The Comparative Statics of Optimal Hierarchies." *American Economic Journal: Microeconomics*, 11(2): 1–25.

Cilliers, Jacobus, and James Habyarimana. 2023. "Tackling Implementation Challenges with Information: Experimental Evidence from a School Governance Reform in Tanzania."

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

Das, Jishnu, Abhijit Chowdhury, Reshmaan Hussam, and Abhijit V Banerjee. 2016. "The impact of training informal health care providers in India: A randomized controlled trial." *Science*, 354(6308): aaf7384.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–96.

De Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2022. "Difference-in-differences estimators of intertemporal treatment effects." National Bureau of Economic Research.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2017. "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia*." *The Quarterly Journal of Economics*, 133(2): 993–1039.

Deserranno, Erika. 2019. "Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda." *American Economic Journal: Applied Economics*, 11(1): 277–317.

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

Deserranno, Erika, Stefano Caria, Philipp Kastrau, and Gianmarco León-Ciliotta. 2022. "The Allocation of Incentives in Multi-Layered Organizations." Northwestern University Working Paper.

Dessein, Wouter. 2002. "Authority and Communication in Organizations." *The Review of Economic Studies*, 69(4): 811–838.

Dhaliwal, Iqbal, and Rema Hanna. 2017. "The devil is in the details: The successes and limitations of bureaucratic reform in India." *Journal of Development Economics*, 124: 1–21.

Dickinson, David, and Marie-Claire Villeval. 2008. "Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories." *Games and Economic behavior*, 63(1): 56–76.

Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review." *The Journal of Human Resources*, 37(4): 696–727.

Duflo, Esther, Rema Hanna, and Stephen P Ryan. 2012. "Incentives work: Getting teachers to come to school." *American Economic Review*, 102(4): 1241–78.

Education Commission. 2023. "Deliberate Disrupters: Can Delivery Approaches Deliver Better Education Outcomes?" *Technical Report*.

Falk, Armin, and Michael Kosfeld. 2006. "The hidden costs of control." *American Economic Review*, 96(5): 1611–1630.

Fenizia, Alessandra. 2022. "Managers and productivity in the public sector." *Econometrica*, 90(3): 1063–1084.

Finan, Frederico, Benjamin A Olken, and Rohini Pande. 2015. "The personnel economics of the state." *Handbook of Economic Field Experiments*.

Finer, S.E. 1997. *The History of Government from the Earliest Times: Volumes I-III*. Oxford University Press, USA.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277.

Heckman, James J, and Jeffrey A Smith. 1999. "The pre-programme earnings dip and the determinants of participation in a social programme. Implications for simple programme evaluation strategies." *The Economic Journal*, 109(457): 313–348.

Hoehn, John R, Caitlin Campbell, and Andrew S Bowen. 2021. "Defense primer: What is command and control." Congressional Research Service. https://crsreports. congress. gov/product ....

Honig, Dan. 2021. "Supportive management practice and intrinsic motivation go together in the public service." *Proceedings of the National Academy of Sciences*, 118(13): e2015124118.

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

Kane, Thomas J, and Douglas O Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives*, 16(4): 91–114.

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

Khan, Muhammad Yasir. 2020. "Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan." *Unpublished manuscript.*

Lang, Kevin. 2010. "Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member." *Journal of Economic Perspectives*, 24(3): 167–82.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. 2021. "Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools." *American Economic Review*, 111(7): 2213–46.

Malik, Rabea, and Faisal Bari. 2022. "Improving service delivery via top-down data-driven accountability: Reform enactment of the Education Road Map in Pakistan." Working Paper.

Mansoor, Zahra, Dana Qarout, Kate Anderson, Celeste Carano, Liah Yecalo-Tecle, Veronika Dvorakova, and Martin J. Williams. 2023. "A Global Mapping of Delivery Approaches." *Technical Report*.

Mehmood, Sultan. 2022. "The impact of Presidential appointment of judges: Montesquieu or the Federalists?" *American Economic Journal: Applied Economics*, 14(4): 411–445.

Muralidharan, Karthik, and Abhijeet Singh. 2020. "Improving Public Sector Management at Scale? Experimental Evidence on School Governance India." National Bureau of Economic Research Working Paper 28129.

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

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy*, 119(1): 39–77.

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

Rambachan, Ashesh, and Jonathan Roth. 2022. "A More Credible Approach to Parallel Trends." Working Paper.

Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal*, 128(608): 413–446.

Rasul, Imran, Daniel Rogger, and Martin J Williams. 2020. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory*, 31(2): 259–277.

Riaño, Juan Felipe. 2021. "Bureaucratic nepotism." *Available at SSRN 3995589*.

School Education Department. 2018. "Annual School Census." Government of Punjab.

Schuster, Christian, Kim Sass Mikkelsen, Daniel Rogger, Francis Fukuyama, Zahid Hasnain, Dinsha Mistree, Jan Meyer-Sahling, Katherine Bersch, and Kerenssa Kay. 2023. "The Global Survey of Public Servants: Evidence from 1,300,000 Public Servants in 1,300 Government Institutions in 23 Countries." *Public Administration Review*, 83(4): 982–993.

Simon, William H. 1983. "Legality, Bureaucracy, and Class in the Welfare System." *The Yale Law Journal*, 92(7): 1198–1269.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives*, 24(3): 97–118.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199.

The History of Government Blog. 2022. "The Art of Delivery: The Prime Minister's Delivery Unit, 2001-2005." *https://history.blog.gov.uk/2022/08/26/the-art-of-delivery-the-prime-ministers-delivery-unit-2001-2005/*, Published on August 26, 2022.

Vivalt, Eva. 2020. "How Much Can We Generalize From Impact Evaluations?" *Journal of the European Economic Association*, 18(6): 3045–3089.

Wilson, J.Q. 1989. *Bureaucracy.* Basic Books.

World Bank. 2020. "Technical Review of the PMIU Data Information System." World Bank Group Technical Report.

World Bank Group. 2018. "World Development Report 2019: LEARNING to Realize Education's Promise." World Bank Publications.

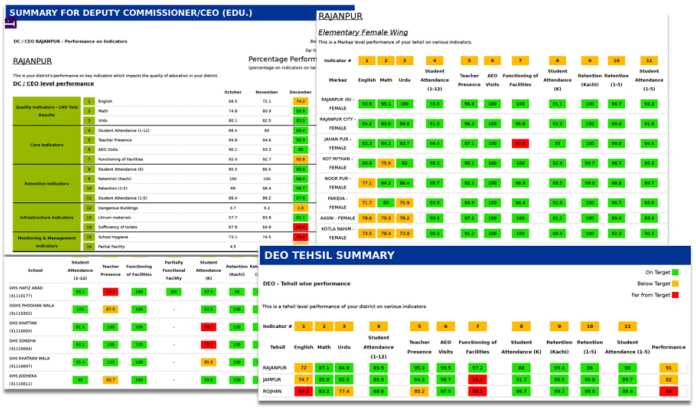
Xu, Guo. 2018. "The costs of patronage: Evidence from the british empire." *American Economic Review*, 108(11): 3170–3198.

Online Appendix

A Data and Design Details

A.1 Data pack

Figure A1: Data pack screenshot



A.2 Color-coded performance thresholds

Teacher presence at every aggregation was coded red when it fell below 86%, orange when it was 86% and above but below 90%, and green when it was 90% or higher. These thresholds were the same for all districts and for all months of the year. Functioning facilities thresholds were 90% and 95%, and were the same across all districts and months of the year.

Thresholds for student attendance varied across districts and months. Districts were divided into three categories, A, B and C, where category A consisted of historically highest performing districts, category C consisted of historically lowest performing districts, and B consisted of the rest. Further, the months in the year were divided into two groups - December-March were considered high attendance months and April-November were considered low attendance months. This division accounted for differential attendance expected due to exams during the school year. Different thresholds were set for each category of districts, for each group of months; for category A districts during December-March, student attendance was coded red if it was

below 89%, orange if it was 89% and above but below 92%, and green if it was 92% and above. During April-November, the thresholds were 87% and 90%. For category B districts the thresholds were 87% and 90% during December-March and 85% and 88% during April-November. For category C districts the thresholds were 84% and 87% during December-March and 82% and 85% during April-November.

A.3 Compliance

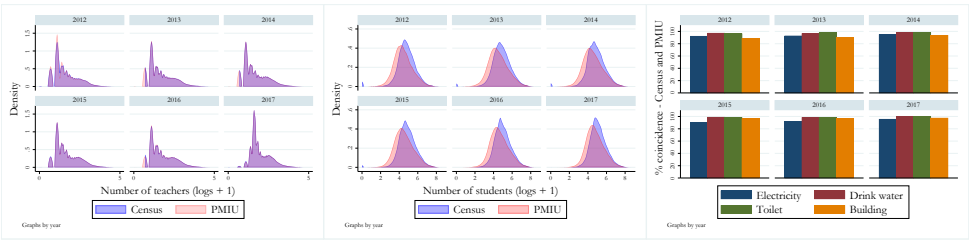
We have data-pack reports for 60 months from December 2011 to May 2018, which account for 100% of the reporting (without June, July, and August). To asses the quality of the data we compare the data-pack reports with the annual school census in the month that the relevant census was undertaken (October). Figure A2 compares the distribution of the variables both sources report. Panels (a) and (b) show for teacher presence and student attendance that both sources overlap, suggesting that the population was mapped consistently. Panel (c) plots the percentage of schools where the functionality coincides, which is near 100%. Panel (d) and (e) plots the distribution of the differences in the reporting. For teacher presence, we observe almost no difference. For enrolled students, it shows a high mass around zero, though a tail of negative values.

A.4 Stacking process

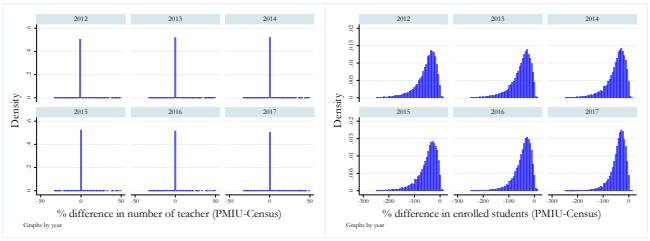
Figure A3 describe the stacking procedure. Each row/column corresponds to a subject/period treatment status. Green indicates treatment. We (arbitrarily) choose one period before and the period of treatment adoption. For S_1, in t_{+1}, units $S_{2,3,4}$ are controls. For S_2 in t_{+2} units $S_{3,4}$ are not treated. For S_3 in t_{+3}, unit S_4 is not treated. For each treated unit we build a two-period panel with its own controls, assign a unique identifier for each, and stack them together by normalizing in relative time so no bias from treatment timing adoption appears from using two-way fixed effects. As the same unit can appear at different events, the fixed effects must be interacted with panel identifiers to account for repeated units and differences in relative time origins.

Figure A2: Data validation - monthly PMIU vs. Census

(a) Overlap - Teacher presence (b) Overlap - Stud attendance (c) Overlap - Funct facilities

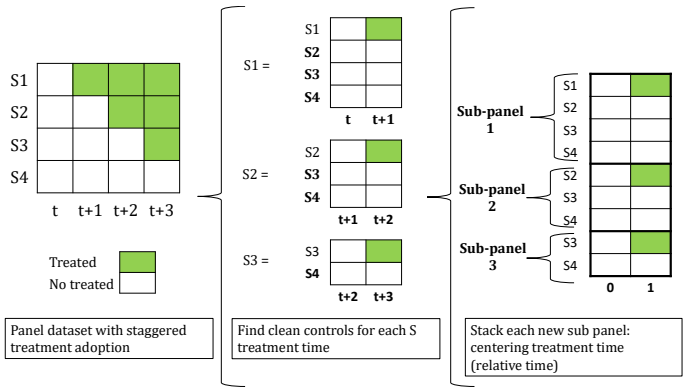


(d) Diff - Teacher presence (e) Diff - Stud attendance



Note: This figure compares October PMIU data and corresponding school-level quantities from the Annual School Census. Panel (a) and (b) plot the distribution of (log+1) teachers and students. Panel (c) plots the coincidence in the reporting of functional facilities (= 1 if functional). Panel (d) and (e) plots the distribution of school differences as percentage change (PMIU - Census)/PMIU dropping the data below percentile 1 and above percentile 99.

Figure A3: Stacking process



A.5 Ranking of district officer positions

District officers can be rewarded/punished in terms of transfers to more/less preferred postings based on performance. To estimate the effect of the oversight scheme on the career trajectory, we collected information on the postings for each senior officer

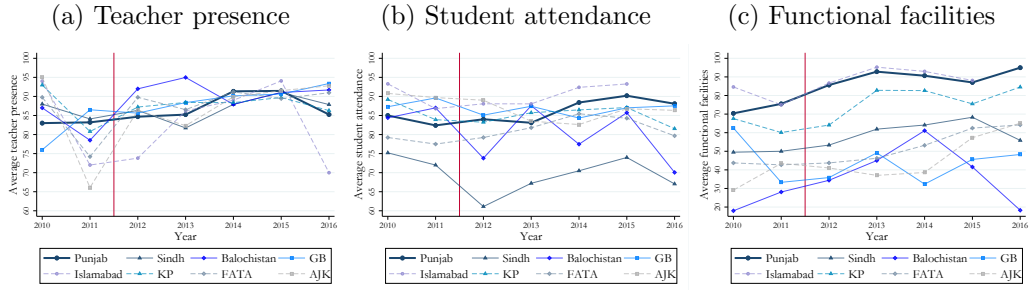
before and after they were posted as district officers. We ranked all designations by seniority to ascertain whether a officer was rewarded/punished determining if a change in position was a promotion/demotion. The ranking of designations was generated through extensive research about seniority levels within the Pakistani bureaucracy and was vetted by two senior bureaucrats.

B Additional Results and Robustness

B.1 Immediate impact of monitoring system implementation

We assess whether the lack of an effect from flagging might be explained by a general impact of the policy across Punjab. Figure B1 shows the average trends of education outcomes in all Pakistan provinces.[20] Note that most provinces are either improving or in a similar trend to Punjab (darker blue line). So despite some underperforming provinces, most of the country faces similar evolving trends.

Figure B1: Pakistan provinces average outcomes trends



Note: The figure show the Data from ASER Pakistan (aserpakistan.org)

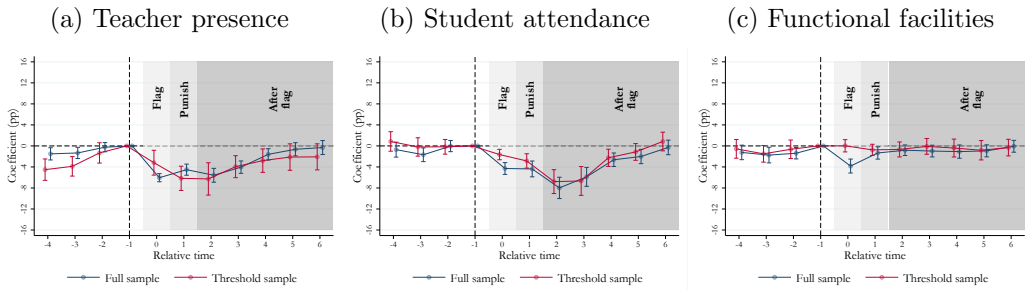
We also test the oversight scheme's immediate impacts by studying the effect of being flagged in the first month of implementation and being a neighbor of a flagged unit. We estimate for the first month of implementation a modified equation 1 including an additional treatment for maraakiz with a school neighboring a flagged markaz. In the equation below $N_{mde} = 1$ represents neighbors, γ_i coefficients for the first time flagged, and β_i for the effect of flagging on neighbors of flagged units.

[20] We recover province-level data for the period 2010-2016 from the Annual Status of Education Report - ASER - Pakistan (aserpakistan.org), which have been independently and consistently conducting household and school surveys to assess the education advancements in the country.

$$Y_{smdte} = \gamma_1(T_{mde} \times Flag_{te}) + \gamma_2(T_{mde} \times Punish_{te}) + \gamma_3(T_{mde} \times AfterFlag_{te}) + \beta_1(N_{mde} \times Flag_{te}) + \beta_2(N_{mde} \times Punish_{te}) + \beta_3(N_{mde} \times AfterFlag_{te}) + \alpha_{mde} + \lambda_{te} + dt + \epsilon_{smdte} \quad (3)$$

Figure B2 reports the results for the first time flagged units in the full (blue) and threshold (red) samples. We observe no positive effects. Instead, both samples of teacher presence and student attendance suggest that flagged units took more time to recover from the negative shock that leads them to be flagged for the first time. Figure B3 reports the results for the neighbors of flagged maraakiz. We observe no positive effects.

Figure B2: Event study - first time flagging effect on performance - flagged units

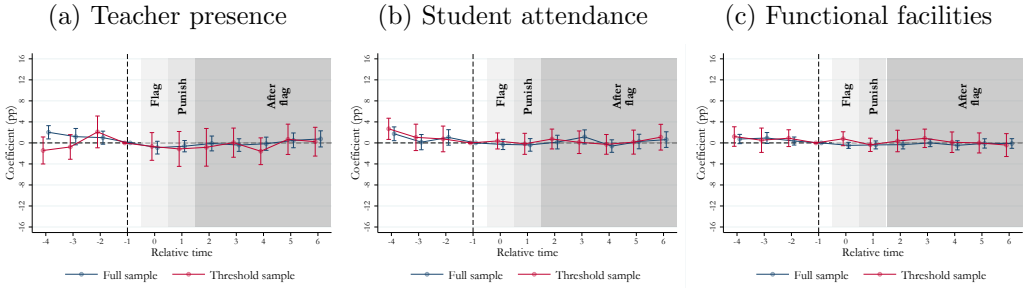


Note: This figure displays the γ_i coefficients from an event study based on equation 3, only for the first month of the oversight scheme implementation, using -1 as the base period, and comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient. Error bars at the 95 percent level are presented for each coefficient.

B.2 Robustness of the stacked design

Because the flagging might turn on/off, the election of post-periods leads us to assume that the unit remains treated, and we might be losing information. We present results to different post-period window stacking, and additional estimators. Figure B4 plots the estimates from estimating equation 1 with a stacking including t periods. *Flag* show the temporary nature of the negative shock. The coefficients of *Punish* remain

Figure B3: Event study - first time flagging effect on performance - neighbor units



Note: This figure displays the β_i coefficients from an event study based on equation 3, only for the first month of the oversight scheme implementation, using -1 as the base period, and comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient. Error bars at the 95 percent level are presented for each coefficient.

qualitatively similar, showing the immediate recovery. The $AfterFlag$ coefficients remain close to zero. Figure B5 reports the event study for a shorter stacking, showing that the trends before and after suggest a similar evolving path as Figure 5. Figure B6 estimates the event study following Sun and Abraham (2021) on the non-stacked dataset, under the assumption of staggered treatment timing, so flagged maraakiz remain treated after the first occurrence.[21] Figure B7 estimates the event study using the DID_l estimator following De Chaisemartin and D'Haultfoeuille (2022), which allows to consider the effect of those switching on/off the treatment, which is also robust to differences in treatment timing.[22]

B.3 Robustness of modeling approach

Table B1 reports the average effect from estimating from equation 1 using flagging history fixed effects, comparing maraakiz that had the same flagging path before the

[21] The authors show that in TWFE dynamic specification with staggered adoption, leads/lag coefficients are contaminated by the effect on other relative periods. It is an special case of Callaway and Sant'Anna (2021) with no covariates (Baker, Larcker and Wang, 2022).

[22] We use a lower number of post periods as the dynamic estimator is obtained as a weighted average of difference in differences comparing the t and $t-l-1$ outcome evolution, between switchers in $t-l$ and non-switchers cohorts (De Chaisemartin and D'Haultfoeuille, 2022).

negative shock. Flagging history is not a markaz attribute, so the term T_{mde} from equation 1 is not absorbed and the interactions can be compared against it. However, by conditioning on past flagging we are generating dependencies in the estimation that make our results more difficult to interpret. Here, we do not have a clean treatment period, and the pre-period now contains some maraakiz that have been flagged, such that the coefficient for T_{mde} is negative for all dependent variables. The average recovery to the mean of these maraakiz then yields a slightly positive coefficient on *After flag* in this specification. However, the net effect of these two coefficients yield qualitatively the same results as in our other tables.

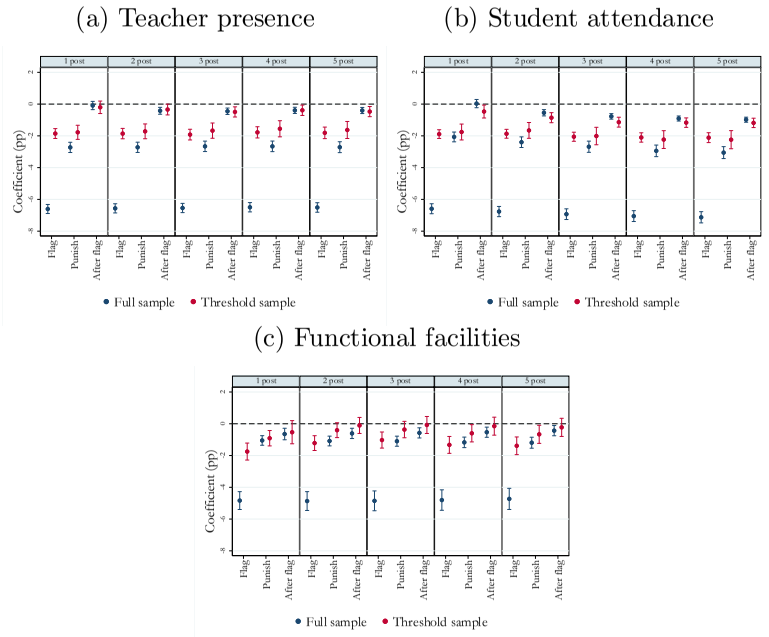
We test for alternative margins of flagging. Figure B8 test the effect of being orange flagged. The results suggest no positive effects after the negative shock. Additionally, Figure B9 test for tehsil level red flagging. The results suggest a non-significant impact of flagging on performance. We also test for changes in the data pack structure involving an increase in the amount of data reported after December 2015 and January 2017. Table B2 show the average results from equation 1, separating by data packs structure, with similar results as the effects on the after-flag period are always closer to zero or non-significant. Finally, Table B3 presents heterogeneity by 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. The results show no differential effect.

Table B1: Monitoring effect on performance - markaz flagging - flagging history FE

Panel A: School outcomes						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T	-2.46***	-0.98***	-4.60***	-2.44***	-8.79***	-2.57***
	(0.097)	(0.10)	(0.15)	(0.14)	(0.27)	(0.12)
T×Flag	-4.47***	-1.11***	-3.45***	0.31**	0.93***	0.69***
	(0.12)	(0.13)	(0.14)	(0.16)	(0.13)	(0.12)
T×Punish	-0.76***	-0.65***	-0.27**	-0.13	2.24***	0.30**
	(0.10)	(0.15)	(0.13)	(0.18)	(0.14)	(0.13)
T×After flag	1.51***	0.35***	2.90***	1.09***	4.83***	1.08***
	(0.11)	(0.12)	(0.17)	(0.16)	(0.22)	(0.12)
N. of obs.	10,331,439	1,870,451	9,414,676	2,192,023	11,636,860	1,994,035
Mean Dep. Var. before	91.4	87.2	88.4	86.3	93.3	91.2
R^2	0.032	0.026	0.17	0.095	0.17	0.039
Panel B: Student scores						
Dependent variable:	Math		English		Urdu	
T	-3.50***	-0.71	-4.18***	-1.96***	-3.72***	-1.33***
	(0.35)	(0.43)	(0.16)	(0.18)	(0.24)	(0.30)
T×Flag	-12.7***	-2.17***	-7.65***	-2.39***	-9.88***	-1.91***
	(0.39)	(0.60)	(0.17)	(0.22)	(0.27)	(0.38)
T×Punish	-1.39***	-0.43	-1.40***	-1.71***	-0.80**	-0.63
	(0.46)	(0.76)	(0.17)	(0.26)	(0.32)	(0.53)
T×After flag	1.80***	0.19	1.79***	0.22	2.06***	0.80**
	(0.35)	(0.58)	(0.18)	(0.21)	(0.24)	(0.37)
N. of obs.	2,281,495	57,196	1,607,728	590,575	2,104,390	150,442
Mean Dep. Var. before	86.6	71.4	74.9	70.3	84.2	71.8
R^2	0.066	0.12	0.060	0.050	0.074	0.11
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

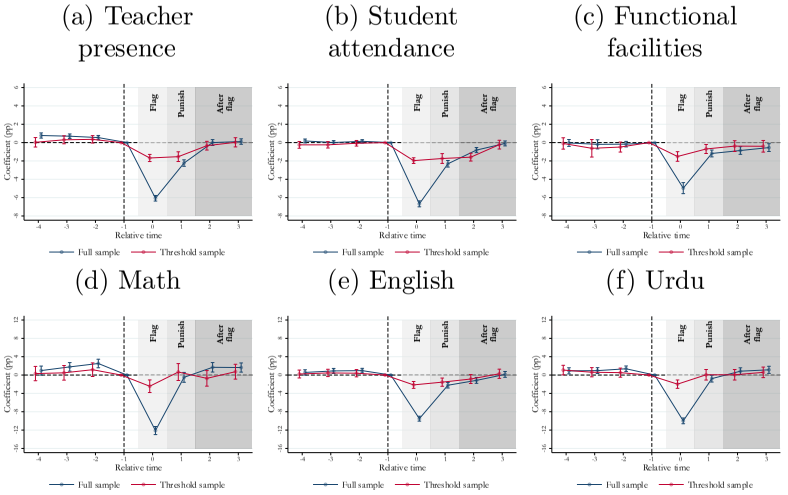
Note: Results from estimating a modified version of equation 1, including flagging history FE instead of markaz FE. Flagging history is built from concatenating the flagging status in the three periods before the observed flagging. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. The flagging and threshold sample are based on the studied outcome. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. 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. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure B4: Average effects by additional after flag t flagging effect on performance



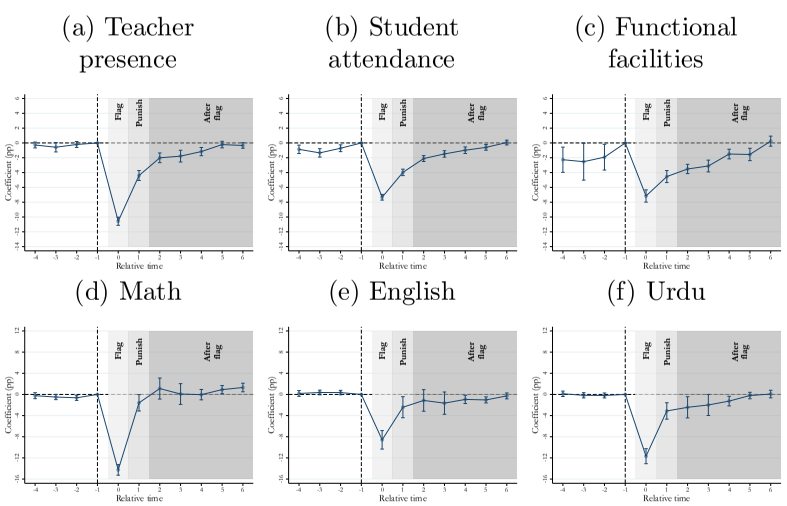
Note: This figure presents results from estimating equation 1 for an stacked dataset including t additional post-periods. The blue coefficients presents results for the full sample, while the red coefficients presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Figure B5: Event study - flagging effect on performance short stack



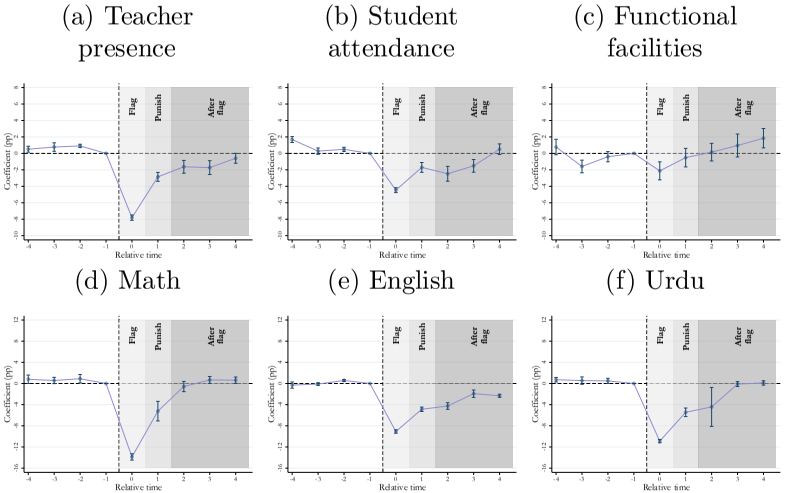
Note: This figure presents results from estimating event studies based on equation 1 using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Figure B6: Alternative specifications
Sun and Abraham (2021)



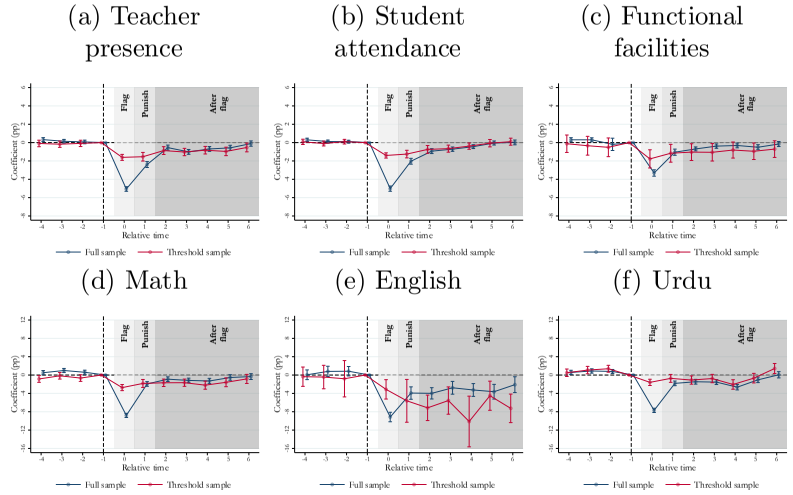
Note: This figure presents the results from estimating an event study based on the Sun and Abraham (2021) difference-in-differences estimator, using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Figure B7: Alternative specifications
DID_l De Chaisemartin and D'Haultfoeuille (2022)



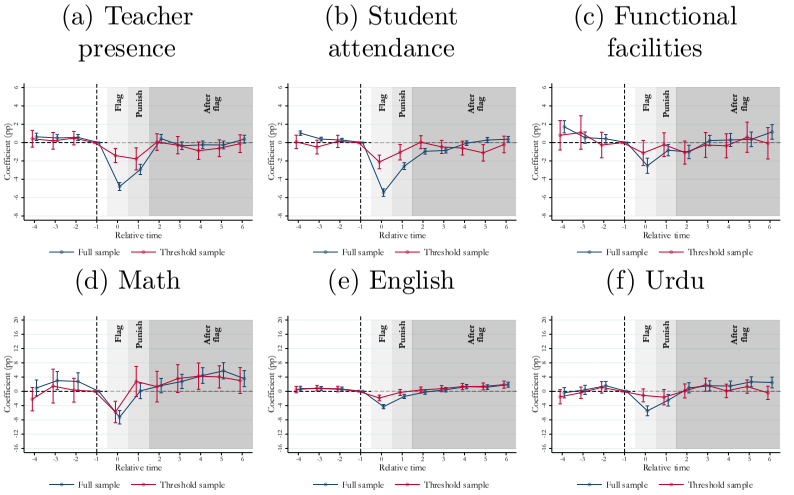
Note: This figure presents the results from estimating an event study based on the DID_l De Chaisemartin and D'Haultfoeuille (2022) difference-in-differences estimator, using -1 as the base period, and three placebo periods before the treatment, comparing schools in flagged and non-flagged maraakiz. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Figure B8: Event study - flagging effect on performance orange threshold



Note: This figure presents results from estimating event studies based on equation 1 using -1 as the base period, comparing schools in orange-flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Figure B9: Event study - flagging effect on performance tehsil-wing flagging



Note: This figure presents results from estimating event studies based on equation 1 using -1 as the base period, comparing schools in flagged and non-flagged tehsil-wing. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Table B2: Monitoring effect on performance by datapack - markaz flagging

Panel A: Datapack 1						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-5.37***	-1.42***	-6.99***	-2.00***	-4.73***	-1.38***
	(0.15)	(0.22)	(0.21)	(0.19)	(0.38)	(0.31)
T×Punish	-2.06***	-1.31***	-3.32***	-2.74***	-1.13***	-0.71**
	(0.15)	(0.26)	(0.22)	(0.36)	(0.19)	(0.32)
T×After flag	-0.69***	-0.55***	-1.19***	-1.62***	-0.54***	-0.37
	(0.12)	(0.20)	(0.10)	(0.21)	(0.18)	(0.32)
N. of obs.	4,960,055	383,685	2,848,511	444,614	4,852,832	350,862
Mean Dep. Var. before	92.3	87.2	91.4	87.5	96.8	94.1
R^2	0.024	0.028	0.10	0.092	0.063	0.072
Panel B: Datapack 2 (After December 2015)						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-6.81***	-2.09***	-7.45***	-2.87***	-3.94***	-0.97
	(0.35)	(0.62)	(0.30)	(0.37)	(0.46)	(0.73)
T×Punish	-1.95***	-1.24*	-0.81***	-0.53	-1.72***	0.11
	(0.50)	(0.73)	(0.23)	(0.35)	(0.65)	(0.81)
T×After flag	-0.11	-0.56	-0.27	-0.073	-2.22**	-0.96
	(0.58)	(1.07)	(0.21)	(0.26)	(1.11)	(1.29)
N. of obs.	971,860	39,343	929,866	66,609	1,138,957	23,832
Mean Dep. Var. before	94.5	87.9	91.8	85.9	98.4	91.9
R^2	0.037	0.046	0.095	0.10	0.068	0.048
Panel C: Datapack 3 (After January 2017)						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-11.2***	-3.84***	-8.00***	-2.38***	-5.74***	-1.67
	(0.41)	(0.42)	(0.42)	(0.50)	(0.62)	(1.35)
T×Punish	-5.50***	-3.46***	-2.02***	-0.90	-2.26***	-1.89
	(0.68)	(1.10)	(0.35)	(0.60)	(0.67)	(1.24)
T×After flag	0.57**	-0.082	-0.65***	-0.53*	-0.81*	-1.29
	(0.26)	(0.34)	(0.21)	(0.32)	(0.49)	(0.91)
N. of obs.	1,047,640	67,922	1,186,456	51,438	1,322,816	17,358
Mean Dep. Var. before	94.2	88.0	93.1	85.9	98.7	92.5
R^2	0.074	0.076	0.11	0.14	0.065	0.054
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

Notes: Results from estimating equation 1 separating by datapack, only for school outcomes due to a lack of scores data for datapacks 1 and 2. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. The flagging and threshold sample are based on the studied outcome. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. 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. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3: Monitoring effect on performance - district ranking and markaz flagging

Panel A: Bottom districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Bottom×Flag×Meeting	-0.13	-0.61	-0.062	0.67	-0.47	-0.55
	(0.49)	(0.55)	(1.29)	(1.28)	(0.72)	(0.83)
Bottom×Meeting	0.56**	0.60**	1.01	1.13*	0.28	0.32
	(0.26)	(0.29)	(0.88)	(0.62)	(0.40)	(0.55)
Flag×Meeting	-4.92***	-4.14***	-4.92***	-5.50***	-0.85	-0.53
	(0.16)	(0.43)	(0.62)	(0.98)	(0.59)	(0.52)
Bottom×Flag×After meeting	0.79	0.31	-1.33	-0.48	-0.41	-0.051
	(0.54)	(0.65)	(1.00)	(1.09)	(0.68)	(0.71)
Bottom×After meeting	0.24	0.20	1.14*	1.49**	0.56	0.27
	(0.30)	(0.37)	(0.58)	(0.55)	(0.54)	(0.50)
Flag×After meeting	-0.39*	0.099	0.78***	0.29	1.25***	1.05**
	(0.20)	(0.48)	(0.20)	(0.59)	(0.26)	(0.47)
N. of obs.	3,063,835	583,417	3,063,410	583,248	3,009,844	565,920
Mean 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	0.47	0.29	-0.36	-3.17	1.25	3.18
	(0.53)	(0.91)	(1.57)	(2.02)	(0.91)	(3.22)
Top×Meeting	0.48	0.33	-1.28**	-0.64	-0.0098	-0.49
	(0.30)	(0.30)	(0.56)	(0.60)	(0.33)	(0.85)
Flag×Meeting	-4.79***	-5.22***	-5.27***	-2.96***	-0.76	-3.21
	(0.23)	(0.66)	(0.57)	(1.02)	(0.56)	(3.31)
Top×Flag×After meeting	-0.57	-1.48	1.17	-0.92	1.06	0.79
	(0.55)	(1.06)	(0.88)	(1.07)	(1.28)	(0.90)
Top×After meeting	0.81***	0.93***	-0.15	-0.30	-0.10	0.49
	(0.24)	(0.21)	(0.43)	(0.61)	(0.20)	(0.39)
Flag×After meeting	-0.049	0.49	0.37*	1.58**	1.29***	1.20**
	(0.15)	(0.67)	(0.18)	(0.71)	(0.26)	(0.54)
N. of obs.	3,111,642	682,461	3,111,048	682,369	3,036,557	672,780
Mean 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

Notes: Results from estimating a modified version of equation 2, including the flagging status of markaz in the quarterly meeting as a third interactions term. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including the schools in the five districts closer to the five in the bottom/top. The bottom/top status and threshold sample are based on the aggregate district performance. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. *Bottom* equals 1 for schools in the bottom five districts and Top equals 1 for the schools in the top five districts on the date of the quarterly meeting. *Meeting* equals 1 in the period of the quarterly meeting. *Mean. Dep. Var before* shows the average outcome in the non-top/bottom districts before the meeting occurs. Standard errors clustered by district, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

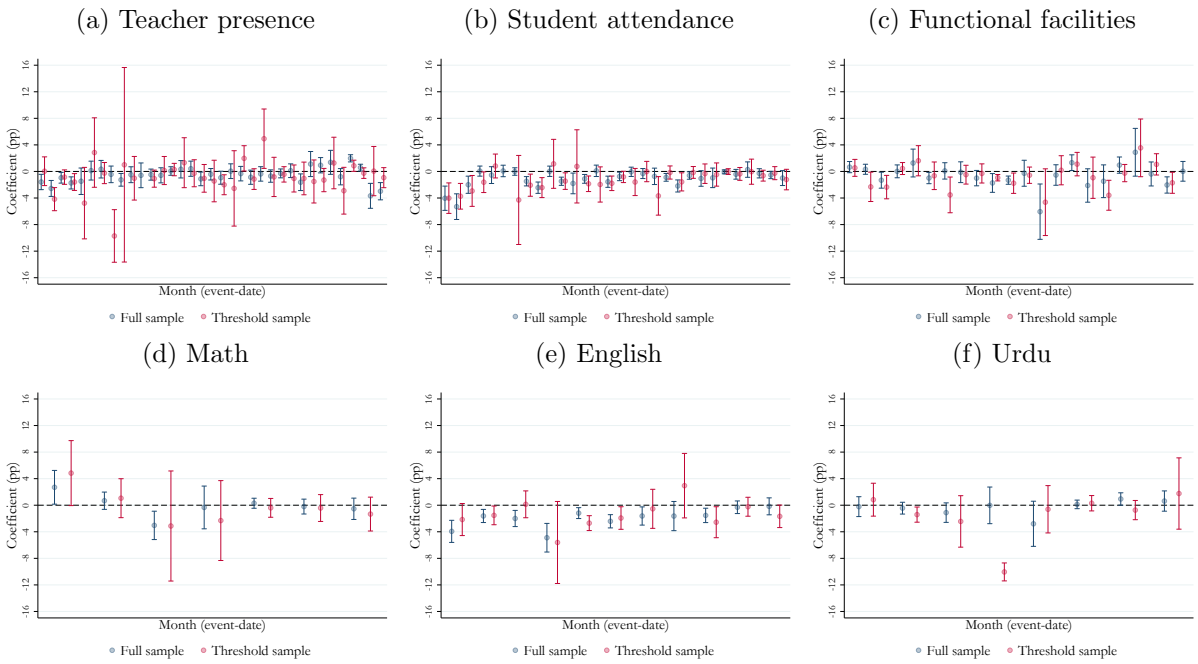
B.4 Other robustness checks

We test additional mechanisms by which results might be confounded. We estimated equation 1 for each specific event to test the robustness of the results to time shocks. Figure B10 reports the coefficients for the $Afterflag$ period. In most of the event panels there appear to be non-significant results, which supports the evidence that on average the centralized monitoring scheme has not improved schools' performance. We tested the reversion to the mean hypothesis by identifying if there existed anticipation of the flagging. The premise follows the assumption that a markaz might start recovering before receiving the flagging if the person in charge knows they might be flagged at the end of the month. We estimated a daily event study where treatment starts once the average outcome of the visited schools on a particular day fell below the flagging threshold. In such a case, we assumed that the public officer might identify the potential flagging and react in the days afterward. The results in Figure B11 suggests no reaction exists in response to being below the threshold for the first time in the month. Finally, we plotted the after-flag β coefficients by accumulating one month at a time an approach to estimate the effect of flagging conditional on the information that the public officer had available for each period. Figure B12 plots the coefficients for each flagging variable. We note that the effect converges towards a null result all cases, and it was possible to identify negative or null effects since the very start of the scheme.

B.5 Assessing the impacts of the scheme across distinct political environments

We assess whether flagging was different in places aligned with the ruling party, considering the pressure set on the scheme by Punjab Chief Minister's participation. We use Provincial Assembly elections data for 2013 to define political alignment as in Callen, Gulzar and Rezaee (2020): an area is politically aligned if the winner of the constituency seat is from the same party as the Chief Minister. We match schools to electoral constituencies to define: i) maraakiz fully aligned: all winners have the same party as the chief minister, ii) not fully aligned, and iii) not aligned: no constituency with the same party as the chief minister. We estimate the following

Figure B10: Seasonality - monthly effects of flagging



Note: This figure presents results from the $AfterFlag$ coefficient by estimating equation 1 for each individual stack (event panel), comparing schools in flagged and non-flagged maraakiz in that particular event. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. 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.

difference-in-differences specification:

$$Y_{mt} = \beta_1(FullyAligned_m \times AfterElection) + \beta_2(NotFullyAligned_m \times AfterElection) + \alpha_m + \lambda_t + dt + \epsilon_{mt} \quad (4)$$

α_m, λ_t are for markaz and time fixed effects, and dt is for district time trends. We limited the analysis to nine months before/after the elections. β_1 β_2 capture the effects of being politically aligned versus not before versus after the election. We also estimate the effect in the sample with high electoral competition, defined by close elections using optimal bandwidth procedures (Calonico, Cattaneo and Farrell, 2020).

Panel A in Table B4 reports the effect on the probability of flagging. Panel B of Table

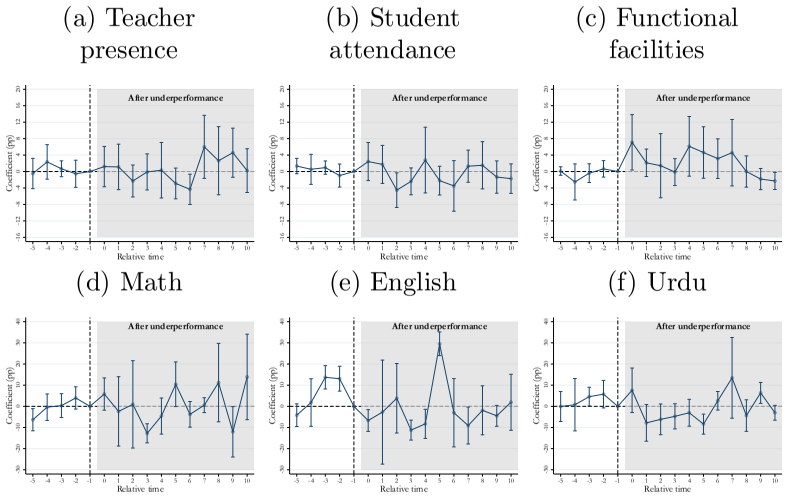
B4 reports the effect on the education outcomes. We do not observe any consistent results across outcomes: there are some effects on student attendance but they do not exist in the close elections sample for the probability of flagging, are very small (under 2 percentage points) for actual student attendance. Thus, we conclude that program implementation was not strongly connected to the ruling party differently from opposition-governed places.

Table B4: Political alignment effect on probability of being flagged

Panel A - Dep. var: Flagging (=1):	Teacher presence		Student attendance		Functional facilities	
Fully aligned×After elections	0.015	0.015	-0.083***	-0.032	-0.014	-0.039
	(0.024)	(0.029)	(0.026)	(0.043)	(0.025)	(0.029)
Not fully aligned ×After elections	0.0024	-0.011	-0.072***	-0.050	-0.014	-0.0077
	(0.024)	(0.029)	(0.026)	(0.037)	(0.024)	(0.027)
N. of obs.	19,143	10,838	18,908	7,804	18,889	10,027
Mean Dep. Var. before	0.086	0.093	0.36	0.35	0.31	0.37
R^2	0.31	0.33	0.37	0.39	0.70	0.72
Panel B - Dep. var: Outcomes	Teacher presence		Student attendance		Functional facilities	
Fully aligned×After elections	-0.29	0.048	1.95***	1.66***	0.12	-0.16
	(0.44)	(0.56)	(0.33)	(0.43)	(0.52)	(0.69)
Not fully aligned×After elections	0.061	0.078	1.43***	1.39***	0.038	-0.67
	(0.44)	(0.58)	(0.32)	(0.38)	(0.57)	(0.77)
N. of obs.	19,141	8,913	19,141	10,277	19,132	9,744
Mean Dep. Var. before	92.0	91.9	86.2	86.0	92.8	92.0
R^2	0.30	0.33	0.54	0.57	0.73	0.77
Sample	Full	Close elections	Full	Close elections	Full	Close elections
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

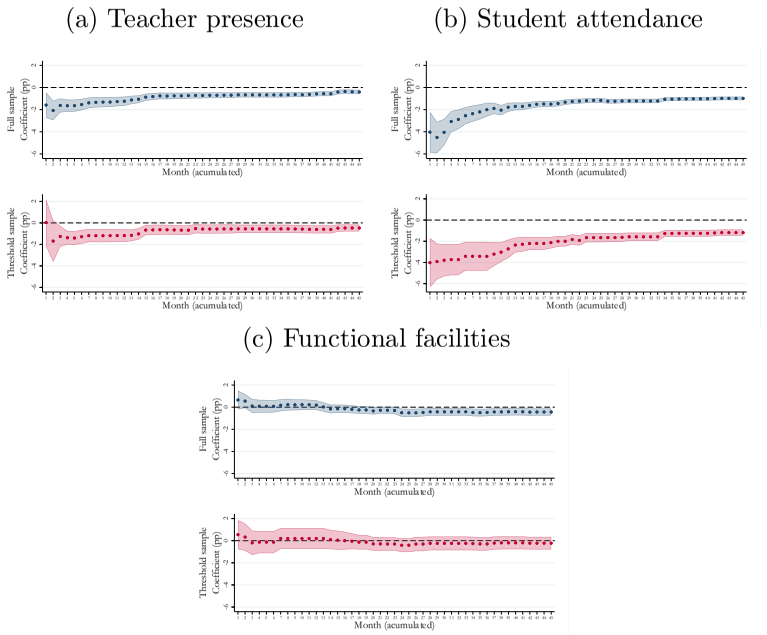
Note: Results from estimating equation 4. Fully (not fully) aligned equals 1 for maraakiz where all (not all) the schools were in a aligned constituency. The control group are maraakiz not aligned. After elections equals 1 for months after May/2013 (elections date). Close elections sample is for the maraakiz with competitive elections, defined as the bandwidth obtained through RD optimization methods around the difference in the vote share between the party aligned and the party not aligned. Panel A report the results on the probability of being flagged for each outcome. Panel B report the results of being politically aligned on the value of the outcomes. Standard errors, clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure B11: Flagging anticipation



Note: This figure presents results from estimating an event study for the daily average of the outcomes in the month of flagging, comparing maraakiz whose average of visited schools was below the threshold at some point in the month, against maraakiz that never underperformed. The base period consists of the day just before the average of the visited schools until that day lies below the flagging threshold. Thus, *After underperformance* is for the periods after the maraakiz was underperforming on average for the first time in the month of data collection (equivalent to *Flag* period in the main specification). Error bars at the 95% level are presented for each coefficient.

Figure B12: Accumulated flagging effects by month after flag accumulated effect



Note: This figure presents results from the $AfterFlag$ coefficient by estimating equation 1 , accumulating one stack (event panel) at a time, comparing schools in flagged and non-flagged maraakiz. The specification accumulates all the months up until t. The blue coefficients presents results for the full sample, while the red coefficients presents results for the threshold sample, obtained through regression discontinuity optimization methods. 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.

B.6 Impacts on the machinery of the government

Table B5: Monitoring effect on other outcomes - effort as mechanism

Panel A: School outcomes flagging						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-0.0037	-0.012	-0.0031	-0.0036	-0.0085	-0.028*
	(0.0049)	(0.0092)	(0.0027)	(0.0052)	(0.0084)	(0.015)
T×Punish	-0.0013	-0.017*	-0.0012	-0.0056	0.013	-0.026
	(0.0058)	(0.0093)	(0.0036)	(0.0073)	(0.0088)	(0.016)
T×After flag	0.00034	-0.015**	0.0052*	0.0059	0.0054	-0.020
	(0.0042)	(0.0073)	(0.0032)	(0.0047)	(0.0062)	(0.013)
N. of obs.	6,208,175	436,428	4,400,630	501,589	6,588,404	356,783
Mean Dep. Var. before	0.97	0.92	0.97	0.95	0.97	0.95
R^2	0.16	0.26	0.20	0.24	0.18	0.25
Panel B: School scores flagging						
Flagging variable:	Math		English		Urdu	
T×Flag	0.011**	-0.0052	0.0067**	0.0040	0.0076**	0.0039
	(0.0044)	(0.0083)	(0.0029)	(0.0056)	(0.0037)	(0.0054)
T×Punish	0.021**	0.017	0.0064*	-0.00068	0.017**	0.0066
	(0.0098)	(0.020)	(0.0035)	(0.0061)	(0.0074)	(0.011)
T×After flag	0.00039	0.0098	0.000072	0.0016	-0.0039	-0.0043
	(0.0036)	(0.0073)	(0.0027)	(0.0044)	(0.0032)	(0.0049)
N. of obs.	1,922,793	45,750	706,106	128,094	1,705,174	103,434
Mean Dep. Var. before	0.98	0.97	0.98	0.97	0.98	0.97
R^2	0.24	0.25	0.23	0.27	0.23	0.30
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: Results from estimating equation 1. The dependent variable equals 1 if the schools received a visit by an AEO. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. 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. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B6: Monitoring effect on other outcomes - gaming in bureaucratic visits

Panel A: Bigger schools						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-0.0052	-0.012	-0.0030	-0.0035	-0.0098	-0.023
	(0.0049)	(0.0093)	(0.0031)	(0.0060)	(0.0090)	(0.015)
T×Punish	-0.0015	-0.018*	-0.00024	-0.0029	0.014	-0.019
	(0.0058)	(0.0096)	(0.0038)	(0.0078)	(0.0096)	(0.016)
T×After flag	0.00066	-0.014*	0.0071**	0.0070	0.0043	-0.014
	(0.0042)	(0.0076)	(0.0032)	(0.0046)	(0.0064)	(0.013)
N. of obs.	3,296,879	239,246	2,339,012	267,953	3,480,301	193,888
Mean Dep. Var. before	0.97	0.92	0.97	0.95	0.97	0.95
R^2	0.19	0.26	0.23	0.24	0.20	0.25
Panel B: Worst performing schools						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-0.00058	-0.012	-0.0046	-0.0063	-0.019	-0.044**
	(0.0065)	(0.012)	(0.0031)	(0.0059)	(0.013)	(0.020)
T×Punish	0.0011	-0.011	-0.00041	-0.0011	0.0051	-0.042*
	(0.0077)	(0.012)	(0.0040)	(0.0080)	(0.014)	(0.024)
T×After flag	0.00063	-0.015*	0.0031	0.0034	-0.0025	-0.033*
	(0.0053)	(0.0091)	(0.0035)	(0.0057)	(0.0092)	(0.018)
N. of obs.	1,419,103	111,744	1,957,278	224,643	546,530	54,055
Mean Dep. Var. before	0.97	0.92	0.97	0.95	0.96	0.94
R^2	0.19	0.28	0.21	0.24	0.21	0.30
Panel C: Schools with most missing teachers						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-0.00058	-0.012	-0.0041	-0.0098	-0.024**	-0.049**
	(0.0065)	(0.012)	(0.0037)	(0.0073)	(0.011)	(0.020)
T×Punish	0.0011	-0.011	0.00095	-0.0010	0.0081	-0.053*
	(0.0077)	(0.012)	(0.0048)	(0.010)	(0.015)	(0.028)
T×After flag	0.00063	-0.015*	0.0059	0.0036	0.0010	-0.030
	(0.0053)	(0.0091)	(0.0041)	(0.0064)	(0.0085)	(0.021)
N. of obs.	1,419,103	111,744	975,682	126,589	1,707,459	98,760
Mean Dep. Var. before	0.97	0.92	0.97	0.95	0.96	0.94
R^2	0.19	0.28	0.25	0.25	0.21	0.28
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: Results from estimating equation 1 only on school outcomes flagging. The dependent variable equals 1 if the schools received a visit by an AEO. The school is the unit of observation for both panels. Panel A estimates for schools whose total number of students is above the median in the maraakiz . Panel B estimates for schools performing below the median of the maraakiz performance. Panel C estimates for schools whose missing teacher rate is above the median of the missing rates in the marakiz. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. 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. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B7: Monitoring effect on yearly school budget

Panel A: OLS estimates				
Dependent variable (in logs):	Total funds	Government funds	Non government funds	Total expenses
Num times flagged (teacher presence)	-0.0057	-0.0030	0.064***	-0.034
	(0.031)	(0.035)	(0.023)	(0.021)
N. of obs.	162,753	162,753	162,753	162,752
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
R^2	0.63	0.63	0.37	0.17
Num times flagged (student attendance)	0.030	0.071**	-0.021	0.025*
	(0.026)	(0.028)	(0.018)	(0.015)
N. of obs.	162,753	162,753	162,753	162,752
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
R^2	0.63	0.63	0.37	0.17
Num times flagged (functional facilities)	-0.0095	-0.012	0.029**	-0.0086
	(0.017)	(0.020)	(0.013)	(0.013)
N. of obs.	162,753	162,753	162,753	162,752
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
R^2	0.63	0.63	0.37	0.17
Panel B: IV estimates				
Dependent variable (in logs):	Total funds	Government funds	Non government funds	Total expenses
Num times flagged (teacher presence)	-0.10	-0.17	0.096	-0.12
	(0.17)	(0.19)	(0.11)	(0.094)
N. of obs.	162,753	162,753	162,753	162,752
F-stat	136.0	136.0	136.0	136.1
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
Num times flagged (student attendance)	-0.37	-0.46	0.082	0.053
	(0.55)	(0.64)	(0.33)	(0.28)
N. of obs.	162,753	162,753	162,753	162,752
F-stat	13.4	13.3	13.4	13.4
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
Num times flagged (functional facilities)	-0.94	-0.84	-0.63	0.39
	(1.04)	(1.08)	(0.69)	(0.51)
N. of obs.	162,753	162,753	162,753	162,752
F-stat	2.99	2.99	2.99	2.99
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
Markaz FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes

Notes: Results from estimating a linear regression of the number of times flagged by each variable in the previous fiscal year on the budget distribution of the current fiscal year. Log-transformed dependent variable. Total funds is the sum of Government funds and Non-government funds. The first report the additional resources transferred from the government. The later report any other additional financial resources transferred to the school. The explanatory variable is the number of times a school was flagged in the previous fiscal year. The regressions include markaz fixed effects, year fixed effects, and district time-trends. Panel B report the second stage from an IV regression where we instrument the number of times flagged by the distance to the flagging threshold, which is akin to fuzzy RD setup. The F-statistic comes from the first stage of the regression to test for weak instruments. Standard errors are clustered at the markaz level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B8: Monitoring effect on labor market - change of head teachers

Panel A: School outcomes flagging						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	0.0047*	0.0022	-0.0050**	-0.00011	0.0059	0.0098
	(0.0025)	(0.0042)	(0.0024)	(0.0050)	(0.0057)	(0.0091)
T×Punish	-0.00099	0.00059	-0.0039	0.0037	0.0013	0.0088
	(0.0026)	(0.0051)	(0.0025)	(0.0051)	(0.0051)	(0.0084)
T×After flag	-0.0015	-0.0016	-0.0038*	0.0022	0.0040	0.0070
	(0.0023)	(0.0038)	(0.0020)	(0.0037)	(0.0030)	(0.0058)
N. of obs.	6,979,870	490,971	4,965,559	562,830	7,409,613	398,961
Mean Dep. Var. before	0.060	0.058	0.059	0.056	0.057	0.069
R^2	0.094	0.10	0.099	0.097	0.084	0.080
Panel B: School scores flagging						
Flagging variable:	Math		English		Urdu	
T×Flag	-0.013	-0.028	-0.0050	-0.00022	0.00066	0.0057
	(0.0085)	(0.017)	(0.0051)	(0.010)	(0.0064)	(0.0091)
T×Punish	-0.025***	-0.016	-0.015***	-0.016	-0.019***	-0.017*
	(0.0073)	(0.013)	(0.0048)	(0.011)	(0.0050)	(0.0087)
T×After flag	-0.014***	-0.024***	-0.0078**	-0.013***	-0.0099**	-0.0032
	(0.0050)	(0.0091)	(0.0033)	(0.0049)	(0.0041)	(0.0060)
N. of obs.	2,198,503	53,361	810,829	147,630	1,950,809	119,592
Mean Dep. Var. before	0.061	0.093	0.071	0.090	0.061	0.063
R^2	0.13	0.13	0.13	0.15	0.13	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: Results from estimating equation 1. The dependent variable equals 1 if the head teacher is different from the one reported in $t-1$. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. 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. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B9: Monitoring effect on labor markets - change of district officers

Dependent variable:	Change of DC			
	(1)	(2)	(3)	(4)
Bottom×Meeting	0.038	0.048		
	(0.048)	(0.062)		
Bottom×After meeting	-0.0075	0.0073		
	(0.026)	(0.037)		
Top×Meeting			0.037	0.075
			(0.058)	(0.051)
Top×After meeting			-0.0052	0.029
			(0.026)	(0.037)
N. of obs.	2,921	605	3,025	685
Mean Dep. Var. before	0.060	0.066	0.058	0.047
R^2	0.15	0.24	0.14	0.31
Flagging	Bottom	Bottom	Top	Top
Sample	Full	Threshold	Full	Threshold
District FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes

Notes: Results from estimating equation 2. The district is the unit of observation for both panels. The dependent variable equals 1 if the district commissioner is different from the one reported in $t-1$ The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including the five districts closer to the five in the bottom/top. *Bottom* equals 1 for schools in the bottom five districts and Top equals 1 for the schools in the top five districts on the date of the quarterly meeting. *Meeting* equals 1 in the period of the quarterly meeting. *Mean. Dep. Var before* shows the average outcome in the non-top/bottom districts before the meeting occurs. Standard errors clustered by district, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B10: Monitoring effect on labor markets - current position of district officers

Dependent variable:	Rank of current employment	
	(1)	(2)
Months in bottom districts	-1.21	
	(1.24)	
Months in top districts		0.39
		(1.14)
N. of obs.	82	81
Mean Dep. Var. before	2.74	2.79
R^2	0.010	0.0011

Notes: Results from estimating a linear regression of the number of months a district officer was ranked in the top/bottom in the quarterly meetings during its time in office on the rank of the current employment. Higher value in the rank account for better career trajectory for district officer. For details on the construction of the rank of employment variable see Appendix A.5. The data is aggregated as the district officer level. Bootstrapped standard errors in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.