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

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

October 10, 2025

A long-established approach to management in government has been the transmission of information up a hierarchy, centralized decision-making, and centralized accountability; colloquially known as 'command-and-control.' This paper examines the effectiveness of a centralized accountability system implemented at scale in Punjab, Pakistan. The scheme automatically identified poorly performing jurisdictions for the attention of central management. Using random variation in the intensity of the scheme's implementation, we show that corresponding de facto punishments had negligible impact on school or student outcomes. We use detailed data on key elements of the education production function to show that a fundamental component of command-and-control approaches to managing the general public sector does not induce bureaucratic action towards improvements in government performance. Using naive before-after comparisons as policymakers' prior beliefs on the impacts of the program, we use a Bayesian framework to show that the posterior distribution lies very close to our estimated effects near zero. With the belief that the scheme was working, the pressures applied by senior management to lower levels of public administration were misaligned with the true areas of need and did not effectively motivate action towards strengthening the education system.

**Keywords**: Accountability, Bureaucracy, Education, Government

**JEL codes**: D73, H11, H83

*We gratefully acknowledge funding from the Blavatnik School of Government/Education Commission DeliverEd program, and the World Bank's i2i initiative, Knowledge Change Program, and Governance Global Practice. We thank Belen Torino for excellent research assistance and the Punjab Program Monitoring and Implementation Unit for providing us with data and institutional details. 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 Political Science and Keough School of Global Affairs, University of Notre Dame; Ladino: Department of Economics, Stockholm University; Mehmood: Independent Scholar; Rogger: World Bank Development Impact Evaluation Research Department.

# 1 Introduction

How the bureaucracy performs is fundamental to the provision of high-quality public services to ordinary citizens around the world (Besley et al., 2022). The vast majority of reforms to government administration implemented at scale relate to shaping the de facto incentives in the bureaucracy instead of introducing changes in legal and fiscal environments.[1] 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.[2] The incentive effects of such schemes depend on de facto accountability to senior managers cascading down the chain of hierarchy.

Governments around the world, faced with constraints on de jure changes in public sector incentives, have been adopting command-and-control management rapidly, with scarce evidence on their efficacy. Following the purported success of British Prime Minister Tony Blair's 'delivery unit',[3] over 80 countries, mimicking this original 'delivery approach', 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). While, what distinguishes these command-and-control approaches in public administration is the remarkable political and executive backing they have received around the world, evidence on these reforms, or the incentive effects of their accountability components, remains thin.

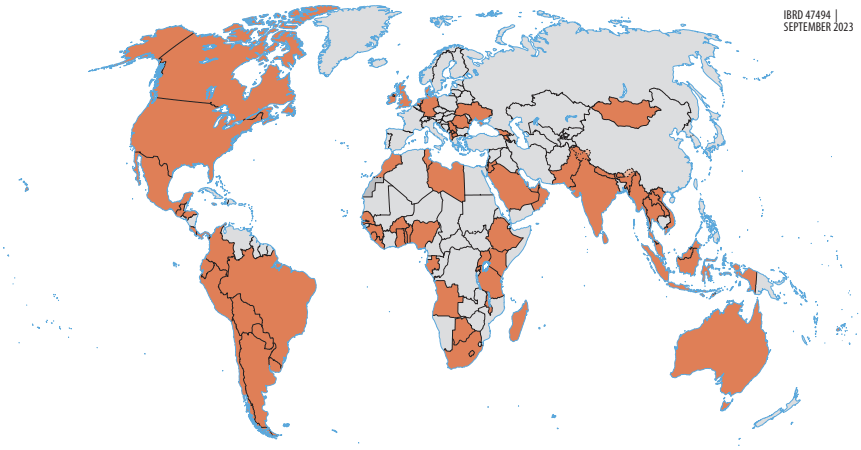
We study the implementation of a system-wide command-and-control delivery approach scheme in the education public administration of Punjab, Pakistan, and focus on the assessment of its accountability effects. Monthly education data from over 50 thousand public schools was channeled to the highest executive authority and used to set targets and establish accountability throughout the organizational hierarchy. 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

[1] 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. Recent research, however, has tended to focus on the important task of evaluating 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).

[2] 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. Many military administrators around the world rely on command-and-control for effective governance across the hierarchy (Wilson, 1989; Hoehn, Campbell and Bowen, 2021).

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

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




*Source:* Mansoor et al. (2023)

the most senior members of the executive (Barber, 2013; Chaudhry and Tajwar, 2021; Malik and Bari, 2023).[4]

Our analysis focuses on how the intensity of accountability implemented within the scheme affects administrative actions and educational outcomes. We collect administrative data from all 52,000 public 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 administrative units.[5] 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 January 2016, student test scores on standardized exams.

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. These additional data allow us to unpack impacts across the hierarchical chain and over a broad range of bureaucratic responses. The scale of the data we have assembled allows

[4] 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 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, 2023). 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."

[5] 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.

us to estimate even small effects with precision, providing an unusually rich picture of the reform.

Using these data, we examine how senior officials' high-frequency monitoring of performance and corresponding efforts to exert control through punishment impact subsequent school performance. We deploy an instrumental variables approach that uses as-if random variation in the intensity of the scheme's execution in Punjab. For each month in our data, the system generates flags on public officials if a sufficient percentage of schools within their jurisdiction (markaz) have fallen below a strict threshold in the educational outcome of interest. We instrument the number of flags a markaz receives in a school year by nine dummy variables for each month of the school year that indicate whether the markaz was flagged and was in a small 'threshold' window around the arbitrary flagging cut-off.[6] The idea is that the annual number of flags, our measure of the intensity of command-and-control accountability, contains a random component that is generated by whether a markaz is 'as-if' randomly flagged within the optimal bandwidth under a local randomization assumption in a given month.[7]

We find precisely estimated evidence that the intensity of accountability in the command-and-control scheme had no substantive impact on targeted school or student outcomes: teacher and student attendance, functional school facilities, as well as English, Mathematics, and Urdu test scores. For instance, our instrumental variables analysis shows that a one standard deviation increase in flagging (reflecting more than a doubling of mean flagging) improves subsequent teacher presence, an outcome clearly within the authority of public managers, by a negligible tenth of a percentage point.[8]

Are there effects simply due to the extensive presence of the scheme? For these results to be consistent with all officials exerting optimal effort due simply to the presence of the scheme, effort levels would have to be unrelated to observed outcomes. A more plausible explanation is that

[6] We also show robustness to more parametric approaches.

[7] The source of variation is similar to the logic of a regression discontinuity design (Cattaneo, Frandsen and Titiunik, 2015; Cattaneo, Jansson and Ma, 2020; Cattaneo and Titiunik, 2022). Many previous studies have similarly estimated causal effects resulting from multiple discontinuity events in the literature on elections and political selection (Folke, 2014; Freier and Odendahl, 2015; Hyytinen et al., 2018; Meriläinen, 2022; Sørensen, 2023; Baskaran, Hessami and Schirner, 2024; Geys, Murdoch and Sørensen, 2024), political representation and development (Clots-Figueras, 2011, 2012; Bhalotra and Clots-Figueras, 2014; Bhalotra et al., 2014; Nellis et al., 2016; Nellis and Siddiqui, 2018; Priyanka, 2020; Bhalotra, Clots-Figueras and Iyer, 2021), and growth and innovation (Campante and Yanagizawa-Drott, 2018; Azoulay et al., 2019; Bahar et al., 2023). We probe the validity of our instrument in a number of ways: we test as-if random assignment by showing precisely estimated balance on lagged school and student outcomes; the distribution of flags is smooth around the cut-offs showing evidence against sorting around the threshold; the instrument produces monotonic variation in the endogenous variable; and, finally, we observe a strong first stage. The exclusion restriction is likely trivially met, since the instrument isolates the variation in a strict subset of the endogenous variable, such that any effects of the instrument on the outcomes should only pass through the endogenous variable.

[8] Point estimates for all six outcomes similarly reflect a precisely estimated change of less than a percentage point, with standard errors smaller than a fifth of a percentage point.

officials simply did not respond to punishments at all. To assess this further, we capitalize on our rich data on administrative activity to study the impacts on key components of the public education production function. We assess the financial and personnel decisions of bureaucratic managers and do not observe more visits from relevant bureaucrats to affected schools, or bureaucratic transfers of head teachers or district heads. We do, however, find changes in budget allocations: maraakiz with a standard deviation more flagging for teacher and student attendance saw an increase of 15 and 18.2 percent of government funding made available to them. However, this amount is small – about sixty additional dollars per annum per school – and there is no evidence that the increase in funds is accompanied by an increase in expenditures related to school operations, consistent with the limited impact on school outcomes. Thus, overall, despite the enthusiasm for the reform of senior managers in Punjab, more command-and-control accountability did not motivate rank-and-file officers to change education outcomes in any substantively significant way.

Despite these null effects, the program was maintained and further developed for six years. A potential reason for the persistence of the program is that a naive examination of before-after comparisons by senior managers yields a strong positive effect of the program. Zooming into monthly data, we find that schools in flagged jurisdictions follow a pattern of return to their equilibrium state of service delivery as their comparison schools in jurisdictions that were not flagged: a reversion to the mean.[9] While senior managers observe the resolution of alert flags via outcome reversion to the mean for particular administrative units, a comparison to an appropriate counterfactual is necessary to conclude that this resolution does not seem to be due to administrator efforts.

Using these reversion to the mean effects as policymakers' prior beliefs on the impacts of the program, we estimate posterior beliefs given the realization of our analysis in a Bayesian framework. We show that the posterior distribution lies very close to our estimated effects, which suggests that across outcomes (teacher presence, student attendance, facilities, and test scores), posteriors concentrate near zero. This Bayesian analysis helps interpret the null effects we document: with the belief that the scheme was working, the pressures applied by senior management to lower levels of public administration were misaligned to the true areas of need and did not effectively motivate action that strengthened the education system.

We contribute to a growing literature on bureaucracy and development broadly (Finan, Olken and Pande, 2015; Besley et al., 2022; Brierley et al., 2023), and on designing optimal incentive structures

[9]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 the public sector more specifically (Ali et al., 2021; Deserranno, Leon and Kastrau, 2022). Recent (frequently experimental) papers in this literature have made the important contribution of showcasing the efficacy of various formal incentive schemes such as performance-related 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), or other non-financial incentives (Ash and MacLeod, 2015; Khan, 2025; Honig, 2021). However, large-scale changes to formal contracting in the public sector have had limited success (Banerjee et al., 2021; Muralidharan and Singh, 2020), and thus bring to the fore the role of de facto incentives.[10]

The de facto incentive most extensively studied in public administration is the role of management (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), and in particular the degree of control a manager attempts to exert over their employees. 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; Craig, Imberman and Perdue, 2015); and less supportive evidence from experiments about administrator's motivation and performance, or those dealing with organizational dynamics (Falk and Kosfeld, 2006; Dickinson and Villeval, 2008; Bandiera et al., 2021; Muralidharan and Singh, 2020). One potential reason for mixed findings is that extensive de jure public service rules and codes have limited explanatory power in the presence of incomplete contracts, a frequently apt characterization of bureaucracies. Consequently, actual (de facto) contracting outcomes in the public sector depend critically on which rules senior managers choose to emphasize or enforce. We provide some of the first at-scale evaluation of de facto accountability and show that corresponding pressure from managers does not engender substantial responses from public officials, however salient senior management makes this form of incentive provision.[11]

[10]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) on how changes in the retirement age impact career concerns in India.

[11]By doing so, our study also 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). The paper provides a lens 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. When taken to scale, such pilots may not provide a sustainable means of managing the public administration (Banerjee, Duflo and Glennerster,

We also add an early contribution to the nascent study of a key feature of bureaucracy: hierarchy. While the theory of hierarchy in organizations continues to develop (Aghion and Tirole, 1997; Dessein, 2002; Chen, 2017; Chen and Suen, 2019; Easterly, 2008), there are few related empirical tests in the literature. Empirical work on government performance broadly finds mixed efficacy of the relative importance of top-down versus bottom-up accountability (Olken, 2007; Björkman and Svensson, 2009; Dunning et al., 2019). Recent evidence from public sector organizations highlights the crucial role of hierarchy in shaping behavior (Deserranno et al., 2022; Cilliers and Habyarimana, 2023). We present the first at-scale evidence in the economics literature on this classic pillar of Weberian bureaucracy: centralized control mechanisms. Given the systemic nature of centralized accountability, command-and-control reforms are poorly suited to experimental evaluation. Rather, our approach extends the use of administrative thresholds as sources of identification on the impacts of top-down accountability as a driver of public sector performance (Chen, Li and Lu, 2018; Bertrand et al., 2020).

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. Section 4 presents the analysis about the effect of the intensity of exposure to command-and-control on schooling outcomes. Section 5 presents assessments of the scheme's impact on key elements of the education administration. Section 6 explores the extent to which the scheme's results respond to naive evaluation of the bureaucratic response. Finally, Section 7 concludes.

## 2 Public Education in Punjab

Home to over half the population of the country, Punjab is the most populous province of Pakistan. Of the 110 million people based there, twenty million are school-aged children, over half of whom attend public schools. With over 52,000 public schools employing 400,000 teachers, the scale of managing public education in the province is substantial (School Education Department, 2018).

The province is divided into 36 districts, which are subdivided into administrative units called tehsils, further subdivided into areas of responsibility called "maraakiz" (plural of "markaz", the Urdu word for "center"). On average, there are four tehsils per district, and 48 maraakiz per tehsil. Thus, on average, a district-level education manager has 192 maraakiz to track, while each markaz-level official manages 20 schools.

The School Education Department is responsible for organizing and overseeing the education sector's performance in the province. The department has two arms: district education authorities,

2008; Banerjee et al., 2021).

which coordinate the implementation of public education delivery, and the Program Monitoring and Implementation Unit (PMIU), which is responsible for collecting and disseminating data on school performance. Both are staffed and organized separately, and monitoring is generally seen as independent of implementation.

### 2.1 Education Implementation and Monitoring

Each district in the province has one district education authority which reports directly to the School Education Department. The district education authority is led by an Executive District Officer (EDO), and three District Education Officers (DEOs).[12] Below the district leadership team, 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 the officers below them. AEOs are the layer of hierarchy above school principals, thus completing a multi-link chain of command from senior executive to school level. Being the lowest level education management official in the district, the AEOs frequent schools more than any other functionary of the education department, and are tasked with working closely with school principals to manage school-level performance (Malik and Bari (2023)).

Such a layered hierarchy is not unusual in administrative settings worldwide, as the physical constraint of traveling to schools, handling administrative tasks for each school, and engaging with head teachers implies a limit on the scale of any officer's ability for oversight. In contrast to this status quo, the promise of command-and-control style large-scale measurement is that it can alleviate physical constraints and centralize the ability to supervise and censure at scale. By dramatically lowering the cost of monitoring individual schools, digitization of public service delivery measurement has opened up the possibility of centralized management throughout the hierarchy. Such a system of monitoring the administration requires an independent administration.

The Program Monitoring and Implementation Unit (PMIU) is tasked with monitoring the performance of district officers. To this end, Monitoring and Evaluations Assistants (MEAs), who report to PMIU, conduct monthly inspection visits of public schools to assess key aspects of the school environment. These monitoring visits are conducted on an unannounced random date every month, and the assignment of school inspections to monitoring assistants is randomized to limit collusion with school staff.

[12] One DEO oversees secondary education (high school and above) directly, while the other two oversee elementary schooling (primary and middle school grades, catering to children aged 4 to 12 years). This paper focuses on the layered organizational structure for management of elementary schools, which constitute 80% of all public schools in the province.

## 2.2 Centralized Oversight Intervention

**Datapacks** Beginning in December 2011, the monitoring data collected by PMIU was used to generate monthly performance reports called 'datapacks.' The datapacks reported performance for key school and student outcomes, including teacher presence, student attendance, visits by education implementation staff, and status of school facilities (electricity, drinking water, toilets, and boundary wall).[13] From 2016, datapacks also reported scores on standardized Math, English, and Urdu tests administered every month.[14]

The reported performance on each dimension was color-coded in the datapacks based on fixed performance thresholds set by the chief minister's team. A jurisdiction could be coded red, orange, or green, with red being the primary flag for underperformance. Figure 4 illustrates the color-coding in the datapacks. Our study period spans from the introduction of the datapacks in December 2011 to May 2018, just before the national elections that led to a change in administration.

**Centralized Accountability using Datapacks** 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.

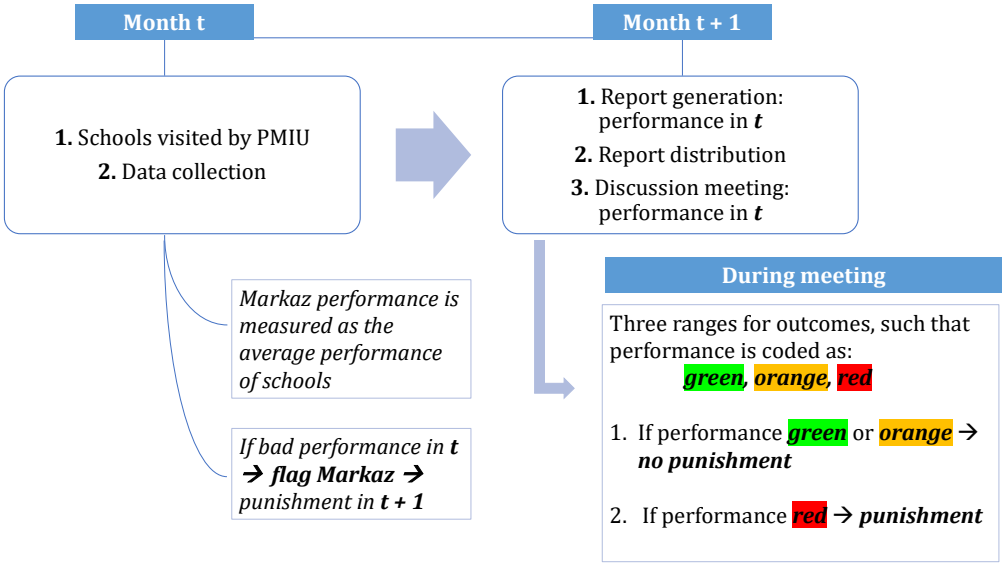
Datapacks were generated to highlight jurisdictional performance starting from the markaz level all the way up to the district level, and a corresponding schedule of meetings was set up where managers probed their subordinates on the reported performance. Figure 2 illustrates the design of the monitoring scheme. Data on all schools in the province was collected in month $t$. Markaz-level average performance was presented to senior district managers in month $t+1$. Maraakiz that did not reach specific (standardized) thresholds were flagged red or orange.

Monthly datapacks at the markaz level, and quarterly datapacks at the district level, were produced from December 2011 to May 2018. These datapacks fed as inputs into accountability meetings at the same frequency with senior managers at the district and provincial levels respectively. This design is a demonstration of centralized, data-driven 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 make it ideal for studying the efficacy of accountability in a

[13] Datapacks 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.

[14] These tests were conducted during the monthly inspection visit by the Monitoring and Evaluation Assistant (MEA). The MEA randomly selected six students from Grade 3 to administer the tests on a tablet through a custom designed testing application.

**Figure 2:** Monitoring scheme structure




command-and-control scheme in the public sector.

## 3 Data

We use administrative data collected at the school level from December 2011 to May 2018.[15] The outcomes are generated from 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 record the status of the school drinking water infrastructure, electricity, toilets, and the boundary wall. We use an aggregate index of the share of functional facilities. Starting in January 2016, PMIU began collecting data on basic literacy and numeracy for grade 3 students through standardized tests covering Math, English, and Urdu, administered by monitoring assistants to seven randomly selected students in each school they visited every month. Scores on these tests for a school are measured as the percentage of correct answers within each of the three subjects. For our main analysis, we construct annual measures of these by taking the average over the school-year. Finally, to understand the effect of bureaucratic behavior, we also use data on district education staff visits to schools.

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

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

posted, enrolled students, and the functionality of school infrastructure. Figure A2 in the Appendix compares both sources and shows that there is a high overlap. A comprehensive review of the data we use assesses it to be of generally high quality (World Bank, 2020).[16]

Table 1 shows descriptive statistics at the markaz-by-year level. There is 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 22, of which nearly 80% are elementary schools. Panel A reports on schooling outcomes. Teacher presence and functional facilities are both above 90% on average. Student attendance, math and urdu scores all show a mean performance above 80%, and english scores have a mean of 77%. Standard deviations range from 5% to 12%, suggesting large variation in outcomes 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. We provide further details about the thresholds for color-coding in Appendix A.

Panel B of Table 1 reports on the intensity of flagging. Maraakiz can be flagged a maximum of nine times a year, once for each month the school is open within the academic year. The mean number of times a markaz is flagged within a year is below one for all outcomes bar english scores (1.191). Table A1 in the appendix reports statistics at the month level by flagging status, and shows the scale of the drop in outcomes associated with flagging varies between 10 and 20 percentage points. 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 and associated flagging was a relatively common feature of the education system in Punjab.

Overall, the relative performance of maraakiz and districts was broadly stagnant. Importantly, a subset of maraakiz remained systematically at the bottom of the distribution. Figure 3 shows the persistence of underperformance. Districts were ranked every quarter based on their overall

[16]The separation of monitoring and implementation spheres of the education administration ensured that any pressures for delivery would not affect the nature of reporting on education outcomes.

**Table 1:** Descriptive statistics - markaz-year

| **Panel A: Schooling outcomes (0-100)** | Mean | Std. Dev. | Min. | Max. | Obs. |
|---|---|---|---|---|---|
| Teacher presence | 92.704 | 5.135 | 0 | 100 | 16,126 |
| Student attendance | 89.115 | 6.927 | 0 | 100 | 16,129 |
| Functional facilities | 92.844 | 12.392 | 0 | 100 | 16,114 |
| Math score | 86.533 | 6.000 | 0 | 100 | 9,482 |
| English score | 77.139 | 7.421 | 0 | 100 | 9,482 |
| Urdu score | 84.884 | 6.003 | 0 | 100 | 9,482 |
| **Panel B: Number of times flagged** | Mean | Std. Dev. | Min. | Max. | Obs. |
| Teacher presence - # times flagged | 0.601 | 1.200 | 0 | 9 | 16,126 |
| Student attendance - # times flagged | 0.963 | 1.801 | 0 | 9 | 16,129 |
| Functional facilities - # times flagged | 0.941 | 2.192 | 0 | 9 | 16,114 |
| Match score - # times flagged | 0.218 | 0.582 | 0 | 9 | 9,482 |
| English score - # times flagged | 1.191 | 1.427 | 0 | 9 | 9,482 |
| Urdu score - # times flagged | 0.266 | 0.639 | 0 | 9 | 9,482 |

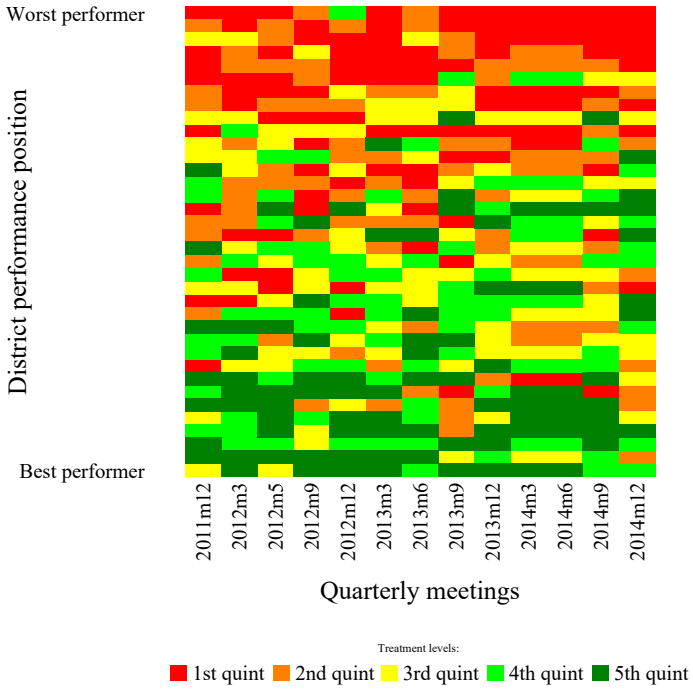
*Notes:* The unit for the variables is markaz-year (specifically, the school year). Panel A reports statistics on the schooling outcomes, measured in percentages from 0 to 100. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students. Number of times flagged in Panel B is for the number of months in the school-year that a markaz was flagged by reporting each respective outcome below the flagging threshold.

performance, so the best (worse) performing districts consist of better (worse) maraakiz. The figure plots for each quarter the quintile in which the district fell in the overall score distribution. It shows that districts in the higher quintiles tend to maintain a high position in the ranking, while districts in the lowest quintiles remain last. The figure thus presents a descriptive sense that the flagging did not motivate poor performers sufficiently for their overall rankings to change. Table A1 in the appendix shows statistics for the change in the ranking for bottom / top-performing districts, further supporting the high persistence of the positions.

The systematic underperformance of some maraakiz is in line with evidence from other settings that indicates that education (and other environments) face structural constraints to improve 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 maraakiz under the flagging threshold irrespective of their baseline performance levels, yet the dominant drivers of service delivery are more structural.

The variation in outcomes between 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, so there is a significant probability of falling

**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.

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.

**Table 2:** Measures of variation by quintile of performance

| **School-level variation (sd) by quintiles of performance** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Outcome (0-100) | Q1 | Q2 | Q3 | Q4 | Q5 | All | Obs. |
| Teacher presence | 10 | .98 | .69 | .75 | 1.4 | 7.6 | 51,532 |
| Student attendance | 14 | 1.3 | .77 | .71 | 1.6 | 9.9 | 51,507 |
| Functional facilities | 17 | 4.3 | 1.8 | .54 | .48 | 16 | 50,500 |
| Math score | 5.6 | 1.1 | .82 | .79 | 1.9 | 6.3 | 37,537 |
| English score | 6.1 | 1.4 | 1.1 | 1.2 | 3.1 | 8.3 | 37,536 |
| Urdu score | 5.8 | 1.3 | .95 | .93 | 2 | 7.1 | 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. 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.

# 4 Intensity of Exposure to Command-and-Control Accountability

## 4.1 Command-and-Control Scheme Implementation

Before proceeding to an analysis of how accountability in the command-and-control scheme shapes education outcomes, we provide details on the high fidelity with which the scheme was implemented. This is not surprising because, as we note in the introduction, what separates these types of reforms is the strong political and executive backing they receive.

**Strong Commitment by Senior Leadership** At the central level, the scheme included quarterly (sometimes bi-monthly) meetings, which were chaired by the Chief Minister, who "would make it a point to not miss any one of the meetings" (Malik and Bari, 2023, p. 32).[17] The senior management of the province placed substantial weight on the system, and the chief minister "had full ownership of this reform, and his presence signaled to the bureaucracy that civil servants were to take the initiatives seriously" (Malik and Bari, 2023, p. 31). Overall, 39 meetings were chaired by the Chief Minister between 2012 and 2018, corresponding to all possible meetings expected for the scheme (Malik and Bari, 2023, p. 31).

Detailed qualitative work by Malik and Bari (2023) describes how accountability worked. Interviews with district officials revealed officers flagged red would get censured in front of their peers: during the meetings "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" (Malik and Bari, 2023, p. 32). Meetings "could result in serious consequences for the districts", and district officials stated that they did "not want to be punished in front of... colleagues" (Malik and Bari, 2023, p. 33). As such, district officers "used to be on their toes because when those top[-performing] three and [the underperforming] three districts were interrogated—[there was a lot of pressure]" (Malik and Bari, 2023, p. 32) [brackets in original]. The political weight and international guidance ensured that accountability protocols were effectively implemented as intended.

This command-and-control accountability approach was replicated inside each district: a senior manager from the district bureaucracy highlighted how the entire district education hierarchy was mobilized in the accountability chain, where the "main actors involved in addressing... problems on the grass root level were the DEOs (District Education Officers) and the Assistant Education

[17] Malik and Bari (2023)[p. 30] 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."

Offices (AEOs) ... These are officers who are in touch with schools at the markaz and individual school level."[18]

**Effective Rollout Across the Province** In each district, datapack reports were used for senior management check-ins within the first ten days of every calendar month. The political signaling by the provincial leadership resulted in near complete compliance of the monthly meetings (Malik and Bari, 2023). For instance, an AEO said the following on his communication with managers: "We meet our relevant DDEO frequently in the office. The DDEO asks us for a daily report of the school situation. We submit the visit plan to the DDEO."[19][20]

Figure 4 illustrates an example of a report used in the monthly meetings. Panel (a) shows maraakiz compared in various performance indicators (in columns). In this example, the English test scores in 'Kot Chutta City Male' markaz and 'Basti Malana Male' markaz were highlighted red for falling below the flagging threshold. Panel (b) presents the part of the report that zooms into 'Basti Malana Male' markaz and shows performance from previous months, highlighting that accountability discussions during monthly meetings focused not only on static underperformance but also on its persistence.

Overall, senior managers did not change de jure conditions, such as making AEO salaries conditional on performance.[21] Instead, the system relied primarily on de facto incentives — long the dominant mode of motivation in public administration — with the aim of transmitting authority and discipline down the hierarchy. The downward transmission of shame and humiliation through hierarchical chains is inherently difficult and may attenuate as it moves from higher to lower levels. At the apex, a district head who is publicly shamed before peers experiences a powerful and visceral loss of status. In seeking to reassert authority, the director may attempt to pass this experience on to subordinates, framing their culpability as shared responsibility. Monthly meetings at the district level did include a degree of censuring of AEOs flagged in the datapacks and calls for improvements in their performance (Malik and Bari, 2023). Yet by the time a centralized message reaches further down the hierarchy, it becomes increasingly diluted: subordinates are no longer humiliated before their peers but are instead subject to reprimand from a single superior. This weaker form of censure does not replicate the director's original experience and generates less

[18] We thank Malik and Bari (2023) for sharing some of their unpublished source material with us.

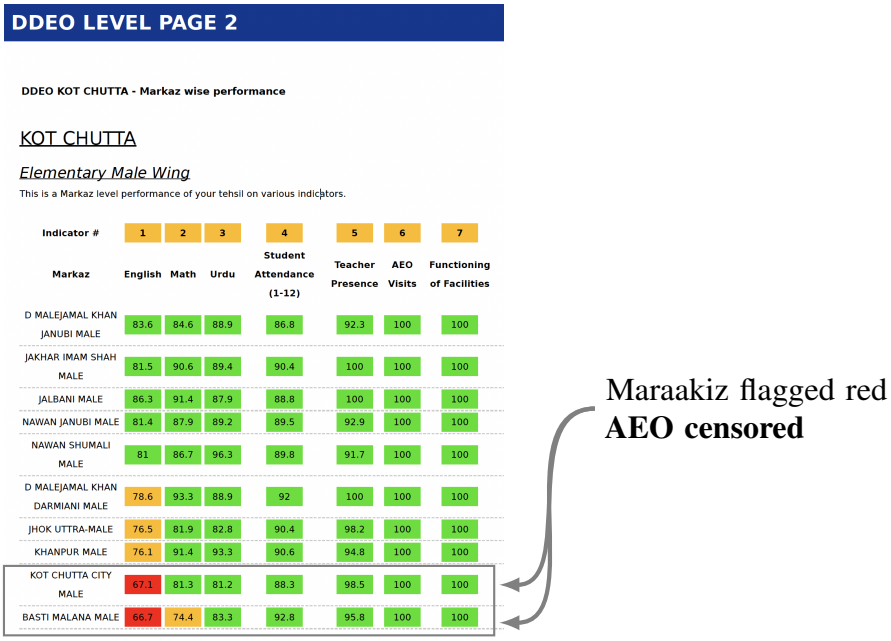
[19] We thank Malik and Bari (2023) for sharing some of their unpublished source material with us.

[20] DDEOs –Deputy District Education Officers– are tehsil (sub-district)-level managers, one administrative level above AEOs, and one level below district managers.

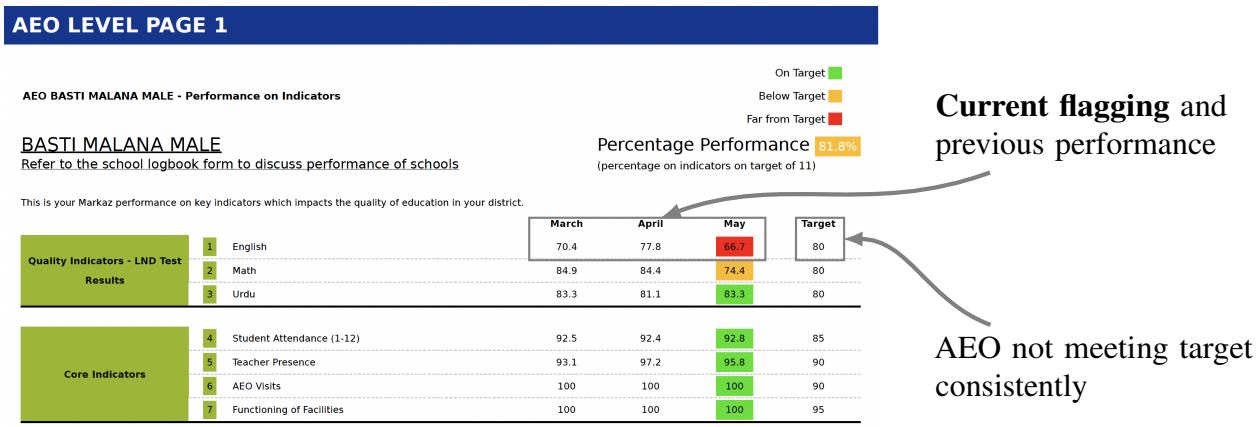
[21] Though some occasional ad hoc financial bonuses were given to district officials. We explore whether there is evidence of staff transfers or long-term impacts on career trajectories from poor performance, but do not find any. Instead, senior management was constrained by public service rules meant to avoid political influence.

**Figure 4:** Datapacks color coding - monthly report for maraakiz flagging and AEO censoring

**(a)** District performance report by maraakiz




**(b)** Details on current and previous maraakiz performance




*Note:* This figure illustrates how AEO/markaz-level performance was presented in the datapacks reviewed during monthly district accountability meetings. Panel (a) displays color-coded performance for each markaz within a jurisdiction (in this case, for maraakiz with male schools in the tehsil of Kot Chutta) across key indicators of interest. Green and orange denote performance exceeding or approaching targets, while red indicates substantial underperformance. AEOs responsible for red-flagged maraakiz were censured in monthly meetings and pressed to address performance lapses. Panel (b) presents a complementary section of the datapack, showing markaz performance over the previous two months, highlighting how monthly accountability meetings focused not just on underperformance in a particular month, but the persistence of underperformance by discussing performance trajectories.

empathy with, or identification in, the leader's sense of shame. The question is whether the process of transmitting de facto 'emotional control' through command structures is an effective channel for enforcing discipline.

### 4.2 Empirical Strategy

To study the impact of flagging accountability in the command-and-control approach on school performance we use variation in the number of times a markaz is flagged during the year. Public officials in charge of maraakiz flagged repeatedly face stronger de-facto punishments, so a more intense exposure to command-and-control accountability should lead them to take actions that induce better school outcomes.

We aggregate our school-by-month data at the school-by-year level so we can count for each markaz the number of times it was flagged during the year. Thus, administrators are more affected by the intensity of flagging if they manage a markaz that was flagged more often. Consequently, the performance of schools in a year depend on the extent to which administrators effectively take actions in response to flagging. We define a school year from September to May of the next calendar year, which corresponds to the first and last month of school activities. The following equation displays the relationship between intensity of flagging accountability and school level outcomes:

$$Y_{s,m,d,t+1} = \beta \cdot TimesFlagged_{m,d,t} + \alpha_m + \lambda_t + \delta_{dt} + \varepsilon_{s,m,d,t+1} \tag{1}$$

$Y_{s,m,d,t+1}$ is the outcome for school $s$ in markaz $m$ and district $d$, for year $t+1$. We measure outcomes as the yearly average performance on teacher presence, student attendance, or functional facilities. $\alpha_m$ denotes markaz fixed effects that control for constant characteristics of maraakiz. $\lambda_t$ is for time-fixed effects to capture year-specific shocks. We include $\delta_{dt}$ –a district binary and linear calendar index– to absorb district linear time trends. $\varepsilon_{s,m,d,t+1}$ is the error term clustered at the markaz level, which is the 'treatment' level. $TimesFlagged_{m,d,t}$ is the z-score of the number of times a markaz was flagged in year $t$, the period preceding the outcome. $\beta$ captures the effect of a standard deviation more intense flagging on subsequent school outcomes.

$\beta$ does not recover a causal effect as the number of times a markaz is flagged can reflect underlying differences not captured in the fixed effects. Therefore, we build an instrument that relies on the discontinuous nature of the flagging that yield random variation in the number of "close-flagging" events to which a markaz is exposed. In a given *year*, a markaz can be flagged up to nine times, one per month of the school-year. For each *month*, a markaz is flagged if it did not pass a performance threshold. If its aggregate performance is barely above the threshold, it is not flagged, despite being very close to a flagged markaz slightly below the threshold. Thus, for each *year*, we can count the number of times a markaz was flagged, and the number of times it was flagged while being close to the threshold.

Under a local randomization assumption (Cattaneo, Frandsen and Titiunik, 2015), flagging in a month is "as good as random" conditional on the markaz performance being within a small bandwidth around the threshold. This produces random variation in each flagging event and allows us to recover exogenous variation from the count of multiple discontinuous events (Borusyak and Hull, 2023). Our approach is similar to those used to estimate causal effects from multiple discontinuity events in the literature on elections and political selection (Folke, 2014; Freier and Odendahl, 2015; Hyytinen et al., 2018; Meriläinen, 2022; Sørensen, 2023; Baskaran, Hessami and Schirner, 2024; Geys, Murdoch and Sørensen, 2024), political representation and development (Clots-Figueras, 2011, 2012; Bhalotra and Clots-Figueras, 2014; Bhalotra et al., 2014; Nellis et al., 2016; Nellis and Siddiqui, 2018; Priyanka, 2020; Bhalotra, Clots-Figueras and Iyer, 2021), and growth and innovation (Campante and Yanagizawa-Drott, 2018; Azoulay et al., 2019; Bahar et al., 2023).

Figure 5 describe the extent of variation by displaying the distribution of the number of times a school was in a flagged markaz on any of the flagging outcomes. The blue bins show the distribution of the total number of times in a flagged markaz, while the red bins indicate the number of times in a flagged markaz that was barely below the flagging threshold.[22] Both distributions show that most of the schools were exposed to a flagged maraakiz at least once and indicate that flagging was a relatively common feature of the education system in Punjab.

### 4.3 Instrumental Variable

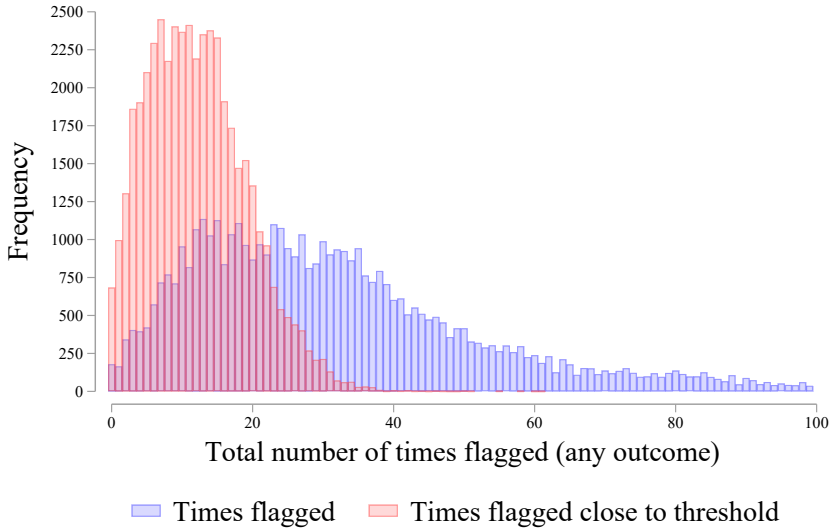
First, we use the threshold definition to identify maraakiz whose average schooling outcomes in a month lie within a small bandwidth on either side of the flagging threshold. We refer to this as the 'threshold sample' in the rest of the paper. We define the optimal bandwidth around the threshold following Cattaneo, Frandsen and Titiunik (2015) and present robustness to it. We obtain the maraakiz in the threshold sample separately for each outcome of interest, and obtain individual bandwidths for each month.[23] Thus, for each schooling outcome, we can count the number of times in the year that a markaz was flagged while being close to the threshold.

Second, we define nine dummy variables, one for each month of the school year, indicating if a markaz was flagged while being close to the threshold, so for each year in our data we can identify

[22]For teacher presence, student attendance and functional facilities, a school can be in a flagged markaz at most nine times in a year. Assuming a school is always in a flagged markaz between 2012 and 2018, it can be exposed to a flagged markaz at most 63 times per outcome, or 189 times in total. The highest number of times flagged by Math, English, and Urdu scores is 22 per score or 66 in total.

[23]The average bandwidth around the threshold for teacher presence is 2.65 pp, for student attendance is 2.73 pp, for functional facilities is 3.25 pp, for Math is 3.19 pp, for English is 4.31 pp, and for Urdu is 3.73 pp.

**Figure 5:** Distribution number of times exposed to a flagged markaz by school




*Note:* This figure illustrates the distribution of the number of times an school was exposed to a flagged markaz. The y-axis reports the number of schools for each bin. The blue distribution shows the number of times a school was ever in a flagged markaz flagged on any of the flagging outcomes. The red distribution shows the number of times a school was in a 'just punished' markaz, i.e. flagged while being 'close' to the threshold of flagging, defined as the optimal bandwidth around the flagging threshold following Cattaneo, Frandsen and Titiunik (2015). The blue distribution is truncated at the $95^{th}$ percentile for illustration purposes.. Table B1 show detailed descriptive statistics at the markaz-by-year level for exposure to the flagging close to the threshold in each schooling outcome.

all the "close-flagging" events. Third, we use the *yearly* panel to instrument the number of times a markaz was flagged in a year with the nine flagging dummies, such that our estimating variation arises from episodes of flagging that were 'as-if' random. We estimate the following first-stage:

$$TimesFlagged_{m,d,t} = \sum_{j=1}^{J=9} \gamma_j \cdot S_{m,d,t(j)} + \theta_{m,d,t} + \alpha_m + \lambda_t + \delta_{dt} + \varepsilon_{m,d,t} \tag{2}$$

where $S_{m,d,t(j)}$ equal to one (zero otherwise) if markaz $m$ was flagged while close to the threshold in month $j$ of year $t$. Using the dummy variables instead of the number of times a markaz was flagged while close to the threshold allows us to measure the intensity of flagging non-parametrically. Thus, we avoid imposing a functional form in the relationship between the endogenous variable and the instrument. We probe this definition in our robustness tests. $\gamma_j$ captures the extent to which being flagged while near the threshold induces additional flagging for a markaz in a year. $\theta_{m,d,t}$ is a time-varying fixed effect equal to the number of times a markaz $m$ was close to the threshold in year $t$. Including $\theta_{m,d,t}$ ensures that the instrument variation comes from episodes where maraaakiz might have been 'randomly' flagged and not from maraakiz whose performance lies far from the threshold. $\alpha_m$, $\lambda_t$, and $\delta_{dt}$ are defined as in equation 1. The following equation defines the second

stage:

$$Y_{s,m,d,t+1} = \beta \cdot \widehat{TimesFlagged}_{m,d,t} + \theta_{m,d,t} + \alpha_m + \lambda_t + \delta_{dt} + \varepsilon_{s,m,d,t+1} \quad (3)$$

Where $\widehat{TimesFlagged}_{m,d,t}$ is the instrumented number of times a markaz is flagged in a year, resulting from the predicted values of estimating the first-stage. Thus, our main parameter of interest is $\beta$, which captures the causal effect of flagging induced by being close to the threshold.

**Identifying Assumptions** To estimate a causal effect we require the instrument to be 'as-if' randomly assigned. We validate this assumption by testing for manipulation of flagging assignment. Figure B1 in the Appendix reports on a test for a discontinuity in the density (Cattaneo, Jansson and Ma, 2020). We observe a smooth distribution around the threshold for all relevant outcomes, except functional facilities. This variable is measured as a share of functional infrastructure among four options, so it is not continuously distributed and peaks in specific values are expected as can be observed in jumps beyond the threshold. For the remaining five outcomes we do not observe significant jumps at the threshold, implying no evidence for manipulation. This is reassuring as the data collection and reporting process makes manipulation unlikely: flagging is based on aggregate performance reports at the markaz level, whose data is collected by randomly assigned agents visiting randomly assigned schools. As such, AEOs have a small capacity to coordinate with all the data collectors and schools in colluding to report better performance.

We further test the random assignment assumption by examining if the treatment variable and instrument correlate with past outcomes. We estimate equation 3 on lagged $(t-1)$ markaz outcomes using as explanatory variables the number of times flagged (D) and number of times flagged while being in the threshold sample (Z). Table 3 shows the results. Panel A reports on markaz schooling outcomes, and Panel B on student scores. The explanatory variables are defined based on the flagging in year $t$ for the outcome reported at the top of the panel. The results come from equation 1, including times close to the threshold fixed effect. Columns (1)-(3) show the results when using the number of times flagged (D) as an explanatory variable. In all cases, there are significant, but smaller than one percentage point differences. In contrast, columns (4)-(6) report the results for the number of times flagged in the threshold sample (Z). We find, in general, an order of magnitude smaller coefficients versus columns (1)-(3). We also observe no significant coefficients, which is consistent with the as-if random assignment of the instrument, so maraakiz randomly exposed to an additional flag are not different on pre-treatment observables from those who were randomly not exposed to it. The only exception is English, where the coefficient is significant but substantively very small.

Taken together, these results show that flagging close to the threshold is plausibly orthogonal to

previous performance, conditional on being close to the flagging threshold.[24]

**Table 3:** Orthogonality of the instrument

**Panel A: Markaz average schooling outcomes**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence$_{t-1}$ | Student attendance$_{t-1}$ | Functional facilities$_{t-1}$ | Teacher presence$_{t-1}$ | Student attendance$_{t-1}$ | Functional facilities$_{t-1}$ |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ – D (z-score) | 0.318** | 0.204*** | -0.747*** | | | |
| | (0.123) | (0.077) | (0.148) | | | |
| # Times flagged threshold$_t$ – Z (z-score) | | | | 0.086 | 0.037 | -0.041 |
| | | | | (0.060) | (0.050) | (0.071) |
| N. of obs. | 7,470 | 7,470 | 7,470 | 7,470 | 7,470 | 7,470 |
| Number markaz | 2,871 | 2,871 | 2,871 | 2,871 | 2,871 | 2,871 |
| Mean Dep. Var | 92.1 | 88.7 | 93.9 | 92.1 | 88.7 | 93.9 |
| Mean # Times flagged | 0.21 | 0.36 | 0.24 | 0.21 | 0.36 | 0.24 |
| SD # Times flagged | 0.59 | 0.82 | 0.77 | 0.59 | 0.82 | 0.77 |

**Panel B: Markaz average student scores**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Math$_{t-1}$ | English$_{t-1}$ | Urdu$_{t-1}$ | Math$_{t-1}$ | English$_{t-1}$ | Urdu$_{t-1}$ |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ – D (z-score) | 0.377*** | 0.591*** | 0.314*** | | | |
| | (0.131) | (0.092) | (0.076) | | | |
| # Times flagged threshold$_t$ – Z (z-score) | | | | -0.026 | 0.208*** | 0.029 |
| | | | | (0.056) | (0.067) | (0.041) |
| N. of obs. | 3,672 | 3,674 | 3,674 | 3,672 | 3,674 | 3,674 |
| Number markaz | 1,836 | 1,837 | 1,837 | 1,836 | 1,837 | 1,837 |
| Mean Dep. Var | 87.1 | 75.8 | 84.2 | 87.1 | 75.8 | 84.2 |
| Mean #Times flagged | 0.032 | 0.28 | 0.050 | 0.032 | 0.28 | 0.050 |
| SD # Times flagged | 0.20 | 0.65 | 0.24 | 0.20 | 0.65 | 0.24 |
| 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 |
| #Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the markaz-year. Results from estimating equation 3. Outcomes in the top of each column, measured in year $t-1$ in scale from 0 to 100. Panel A reports on schooling outcomes. Panel B reports on student scores. # Times flagged$_t$ counts the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ counts the number of times flagged while being close to the flagging threshold in the outcome reported at the top. The variables for the # Times flagged are normalized (z-score). Year is measured as school-year (September to May). Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Standard errors clustered by markaz are in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

We also require the exclusion restriction to hold: the number of times a markaz is flagged close to the threshold (Z) should influence subsequent school performance only through its effect on

[24]This analysis uses 80% of the maraakiz, comprising 94% of schools, observed in the school-year level estimations as maraakiz have changes in names that make the panel unbalanced. We show that our results are robust to estimation on this sample in Table B2 and discuss it below.

the number of times the markaz is actually flagged (D). This assumption is plausible because the thresholds for flagging were set arbitrarily—based on broad, district-wide performance distributions rather than any pre-existing markaz-level cutoffs—and were unrelated to other features of the school system. Moreover, as Malik and Bari (2023) document, accountability discussions were entirely framed around these thresholds. Consequently, the IV is unlikely to affect school performance through any channel other than changes in the number of times a markaz is flagged.

We test for additional IV assumptions in the Appendix. Figure B2 report a positive monotonic relationship between the instrument and the endogenous variable, residualized from the fixed effects reported in equation 2, suggesting that the monotonicity assumption is met. Figure B3 shows evidence for the relevance assumption by reporting point estimates ($\gamma_j$) from equation 2 for each instrument on the number of times flagged (z-score).

### 4.4 Results

Table 4 presents our main results. Panel A reports on school outcomes, and Panel B on student scores. For each outcome at the top of the panel, the first column reports the OLS result and the second column reports the IV result. TimesFlagged$_t$ is defined based on each outcome flagging in year $t$ and standardized such that a unit increase in the treatment is equal to a standard deviation increase in flagging intensity. On average, a standard deviation increase in flagging is equivalent to moving from 0.038 flags in the previous year to 1.61 flags. Outcomes in $t+1$ are scaled from 0 to 100. For all outcomes, the first stage F statistic is high, suggesting that the instruments altogether are not weak.[25]

We find that more intense command-and-control accountability has limited effects. OLS yields small significant effects for all outcomes. For teacher presence, for example, an increase of one standard deviation in flagging increases the average teacher presence by 0.205 percentage points. In Panel A, the 2SLS coefficients indicate an improvement only for teacher presence. The estimate show that a one standard deviation increase in flagging increases teacher presence by 0.112 of a percentage point. This implies that even the 95th percentile of flagging intensity would yield less than a quarter percentage point increase in teacher presence. There are also similar small and negative effects on student attendance and functional facilities, but these are not significant. Perhaps intuitively, teacher presence is the element of the student production function that would be most responsive to hierarchical pressures. But as we report, even in this case the effects are limited. The scale of our data allows us to be relatively precise in estimation, allowing us to detect the extremely

[25] We report the Kleibergen and Paap (2006) Wald test statistic, which coincides with a non-homoskedastic robust F-statistic in settings with a single endogenous regressor (Andrews, Stock and Sun, 2019).

small, and economically negligible, impacts. The results in Panel B also suggest that more intense exposure to accountability does not appear to substantially improve average student performance throughout the year.

**Robustness** We carry out a range of robustness tests to probe our findings. We first use a regression discontinuity aggregation design –RDA– (Borusyak and Kolerman-Shemer, 2024) that resembles our specification with additional controls.[26] Appendix Table B3 shows that the effects remain close to one percentage point.

Second, we probe if the results are sensitive to more parametric definitions of the instrument in Appendix Table B4 and alternative definitions of endogenous variable in Appendix Figure B4. In both cases the coefficients remain smaller than a percentage point. Third, we also test how the effects evolve by year in Appendix Figure B5. There are no clear time trends: teacher presence and student attendance remain flat near zero over time, while functional facilities fluctuates between small positive and negative values. Fourth, we show in Appendix Table B2 that our results are robust to subsetting to the same sample as the orthogonality analysis in Table 3. Fifth, we continue to find null effects when we vary bandwidth sizes around the flagging threshold.

**Alternative Margins of School Response** While we have shown that there is no evidence that more intense flagging improved outcomes on average, it could be the case that the command-and-control accountability worked in either improving the worst or best months of performance in a given school year. We find no evidence for this in Appendix Table B5, which takes as the outcome the performance in the best and worst months of a school year instead of the average across all months. Remarkably, IV estimates show that improvements remain smaller than a percentage point, suggesting that the program did not substantively improve outcomes across the school year.

Next, we examine if, despite the average null, specific schools improved their performance by exploring heterogeneity by school characteristics. We use the fact that whether a markaz is flagged depends on the *average* performance of all schools in that markaz and separate the sample by best- and worst-performing schools, defined as those with an average yearly performance above or below the markaz median. As before, Table B6 reports minimal effects in both types of schools, suggesting that accountability did not move outcomes differentially in best or worst schools in maraakiz.

[26] RDA intend to simulate the local linear estimation of RD designs. In our setting, it consists of yearly averages of the running variable only on the months where the markaz i) was flagged while around the threshold and ii) was close to the threshold. The instrument is the share of months in which a markaz was flagged while around the threshold and additionally control for the share of months in which the markaz was close to the threshold.

**Table 4:** Impacts of exposure to flagging

**Panel A: School outcomes**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ (z-score) | 0.205*** | 0.112* | 0.321*** | -0.067 | 0.946*** | -0.091 |
| | (0.046) | (0.067) | (0.048) | (0.079) | (0.082) | (0.106) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 |
| Mean Dep. Var. | 91.6 | 91.6 | 87.6 | 87.6 | 90.2 | 90.2 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 |
| First stage F-stat | | 210.8 | | 258.6 | | 382.5 |

**Panel B: Student scores**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Math$_{t+1}$ | | English$_{t+1}$ | | Urdu$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ (z-score) | 0.287*** | 0.224 | 0.768*** | 0.580*** | 0.169** | -0.237 |
| | (0.081) | (0.189) | (0.113) | (0.165) | (0.080) | (0.175) |
| N. of obs. | 67,385 | 67,385 | 67,383 | 67,383 | 67,384 | 67,384 |
| Number markaz | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. | 86.9 | 86.9 | 76.5 | 76.5 | 84.6 | 84.6 |
| Mean # Times flagged | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 34.8 | | 157.3 | | 53.3 |
| 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 |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. Results from estimating equation 3. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A reports on schooling outcomes. Panel B reports on student scores. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

Finally, further solidifying our main conclusion that the intensity of accountability did not produce meaningful movement in the bureaucracy, we find null effects when we test if accountability over one outcome produces changes in other outcomes. These results are available on request.

**Change of Incentives for High Performance Thresholds** The negligible accountability effects documented thus far come from the main component of the command-and-control system, in which AEOs were censured when flagged red due to the aggregate underperformance of the schools in their markaz. Yet, it might be that the system produced changes around higher flagging thresholds, where bureaucrats can be rewarded due to good performance. We test in Appendix Table B7 whether there are significant impacts of a more intensive accountability around the green/orange flagging threshold, which we know qualitatively from Malik and Bari (2023) received limited attention from the senior bureaucracy. As before, the results remain smaller than a percentage point for all outcomes, further reinforcing that flagging did not produce strong changes in the bureaucracy.

**Change of Incentives in Higher Levels of Hierarchy** School performance may also be affected based on the district in which they are located, as the district officers are punished (rewarded) if the district is in the bottom (top) five performing districts. In such cases, we might expect command-and-control accountability to have more of a bite because more officers on the administrative hierarchy are primed to act. We explore differential effects by interacting the number of times flagged with an indicator for ever being in the five bottom/top districts in year $t$.[27] We once again find small effects, with the results described in Appendix Table B8, suggesting that additional accountability pressures also did not produce any movement in school outcomes.[28]

Alternatively, effects may appear at more aggregate levels, as district officials could have incentives to raise aggregate performance to remain in the top five – and receive rewards – or to move out of the bottom five – and avoid punishment. We estimate a modified version of equation 3 in which we aggregate the data to the district-by-quarter level and define flagging as being ranked in the top or bottom five. We use as instrument cases in which the district was very close to falling out of the top (bottom) group, and control for cases in which the district is near the corresponding threshold. Appendix Table B9 reports the results. Most coefficients remain below one percentage point or are negative for both bottom and top-ranked districts.

[27] For the 2SLS estimations using 3, we first instrument the number of times flagged using equation 2, then interact the predicted variable with the district ranking indicators.

[28] Data on district quarterly meetings is available until May 2015, before student score data are available. Thus, we only report results for teacher presence, student attendance, and functional facilities.

**Political Pressure on Improved Performance** Another possibility is that the system was intended to serve political ends. We follow Callen, Gulzar and Rezaee (2020), and use data from the 2013 Provincial Assembly elections to explore differential effects by alignment with the state ruling party. We define an aligned markaz if all its schools lie in constituencies with a winner from the Chief Minister's party.[29] Appendix Table B10 shows the results of the interaction of an school alignment indicator with the number of times flagged. IV estimates show no differential effects of alignment with the state ruling party, suggesting that political pressure does not mediate whether the system improves educational outcomes.

## 5 Tracing Impacts Through the Machinery of Government

Although we did not observe meaningful impacts of the accountability component of the command-and-control scheme on school outcomes, the approach to public management may induce a response from bureaucratic actors within the government machinery. We use the detailed data we have assembled to investigate whether we can detect effects along the chain of bureaucratic hierarchy. We are able to analyze impacts on administrative action in terms of both personnel and financial resources, the two key inputs to effective government functioning.

**Bureaucratic Oversight** A natural response by public officials repeatedly flagged for poor performance would be to visit poorly performing schools to undertake diagnostic and remedial work. School visits are a standard part of the AEO work program and a mechanism to resolve issues that schools face in functioning effectively. To measure AEO school visits, we calculate the share of months that a school received a bureaucratic visit in a year. Appendix Table C1 indicates that more intense flagging has no significant effects on bureaucratic visits to schools. Figure C1 shows the robustness of these results for alternative instrument definitions.

**Transfers and Postings** Public officials can also intervene in the management of schools through the labor market by moving head teachers in response to intensive flagging. We explore whether a higher intensity of accountability induced greater rotation of head teachers, as AEOs might use it to improve school performance within their administrative unit. To measure the rotation of head teachers, we define a variable equal to one if the head teacher reported in a month is different from the one reported in the previous month. We then calculate the share of months in a year that a school reported different head teachers. Appendix Table C1 indicates that treatment has no significant

[29] Maraakiz might overlap across constituencies, so we also control for an indicator for not fully aligned.

effects on bureaucratic transfers and postings. Figure C1 further shows the robustness of these results for different instrument definitions.

**Budget Allocation and Utilization** In addition to increasing monitoring intensity, public officials can also channel more budgetary resources to support struggling schools. We estimate equation 3 on measures of school budget allocation and utilization to explore the relationship between command-and-control and school resources.

Our measure of school funds consists of the budget allocated for development spending, which schools use to cover non-recurrent needs. About 86% of this budget comes from the government and the rest comes from non-government sources.[30] While district officials have limited influence over non-government funding, they have substantial influence over government contributions to the development budget. Every year, AEOs assess how much development funding is needed for each school they supervise, and communicate the requirements. If approved, funding is allocated to the district education authority, which has autonomy over its spending.

Table 5 shows the results on the total development funds and its expenditures. With our IV strategy, we find that a one standard deviation increase in flagging for teacher presence and student attendance increase the total amount of funds made available to schools, by 15 and 18.2 percent respectively. However, the increase is minor as for the *average* school this amounts to between 48.3-59.7 additional dollars annually per school.[31] The increase in resources arises solely from government funds. We detect no impacts on non-government funds (see Appendix Table C2). We do not observe a corresponding impact on resource availability arising from more intensive flagging for lapses in functional facilities and student scores variables. This small increase in financial resources is the most direct evidence of the impact of command-and-control management that we detect in this study. Perhaps understandably, it is on the margin of administration senior managers have most influence over.

We find mixed evidence for corresponding changes in expenditure at the school level. In the case of teacher presence flagging, the coefficient suggests that a one standard deviation increase in flagging increases expenditure by 9.2 percent (s.e.=6.5). The coefficient for flagging on student attendance suggests a *decrease* in expenditure by 7.4 percent (s.e.=7.2). Yet, both coefficients are statistically insignificant. Thus, our microdata does not provide evidence that the increase in funding also

[30] Local community participatory bodies are authorized to raise funds for development expenses of school from parents, philanthropists, and other non-government sources.

[31] Between 6,724-8,293 Pakistan rupees, using the average conversion rate from December 2018, corresponding to 139 rupees per dollar.

increased subsequent expenditures related to school functioning, which is in line with the idea that beyond senior management action, command-and-control accountability has limited effects on bureaucratic activity. Figure C2 in the Appendix reports robustness to alternative instrument definitions for these budgetary outcomes.

**Table 5:** Intensity of exposure to flagging - effect on budget

| **Panel A: School outcomes flagging** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Flagging variable | Teacher presence | | Student attendance | | Functional facilities | | Teacher presence | | Student attendance | | Functional facilities | |
| | Dependent variables (logs) | | | | | | | | | | | |
| | Total Funds$_{t+1}$ | | | | | | Expenditure$_{t+1}$ | | | | | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.197*** | 0.147** | 0.149*** | 0.182* | -0.023 | -0.031 | 0.047 | 0.092 | 0.046 | -0.074 | 0.037 | -0.029 |
| | (0.045) | (0.075) | (0.050) | (0.097) | (0.043) | (0.089) | (0.037) | (0.065) | (0.040) | (0.072) | (0.040) | (0.072) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 |
| Mean Dep. Var. (unlogged) | 44,827 | 44,827 | 44,827 | 44,827 | 44,827 | 44,827 | 57,033 | 57,033 | 57,033 | 57,033 | 57,033 | 57,033 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 |
| First stage F-stat | | 210.8 | | 258.6 | | 382.5 | | 210.8 | | 258.6 | | 382.5 |
| **Panel B: Student scores flagging** | | | | | | | | | | | | |
| Flagging variable | Math | | English | | Urdu | | Math | | English | | Urdu | |
| | Dependent variables (logs) | | | | | | | | | | | |
| | Total Funds$_{t+1}$ | | | | | | Expenditure$_{t+1}$ | | | | | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.056** | 0.062 | -0.035 | -0.003 | 0.017 | 0.013 | 0.057** | 0.082 | -0.002 | -0.059 | 0.022 | -0.046 |
| | (0.026) | (0.061) | (0.027) | (0.045) | (0.024) | (0.050) | (0.026) | (0.051) | (0.025) | (0.042) | (0.020) | (0.037) |
| N. of obs. | 67,385 | 67,385 | 67,383 | 67,383 | 67,384 | 67,384 | 67,385 | 67,385 | 67,383 | 67,383 | 67,384 | 67,384 |
| Number markaz | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. (unlogged) | 44,827 | 44,827 | 44,827 | 44,827 | 44,827 | 44,827 | 57,033 | 57,033 | 57,033 | 57,033 | 57,033 | 57,033 |
| Mean # Times flagged | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 34.8 | | 157.3 | | 53.3 | | 34.8 | | 157.3 | | 53.3 |
| Markaz FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| District time trends | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. Results from estimating equation 3. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A report on the schooling outcomes flagging. Panel B reports on the student scores flagging. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Unlogged total funds and expenditure in pakistani rupees. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

Taken together, our results on the machinery of government imply that more intensive command-and-control accountability yields small increases in resources budgeted for schools flagged on teacher and student attendance, but no other actions to convert these resources into improved service delivery. These findings are consistent with our main results implying that the increase in allocated resources had no impacts on educational outcomes.

## 6 Naive Evaluations of Response

Despite the limited impacts documented here, such command-and-control systems are prevalent in many public sector settings. Why do such programs persist? One explanation is that senior managers are employing naive evaluations of the response to such programs. It may be that after arbitrary shocks, the outcomes of interest naturally return to their original state as shown by Chay, McEwan and Urquiola (2005). Yet, if senior managers intervene and do not benchmark treatment jurisdictions with appropriate counterfactuals, they may naively attribute the dynamics of the improved outcomes to their own interventions.

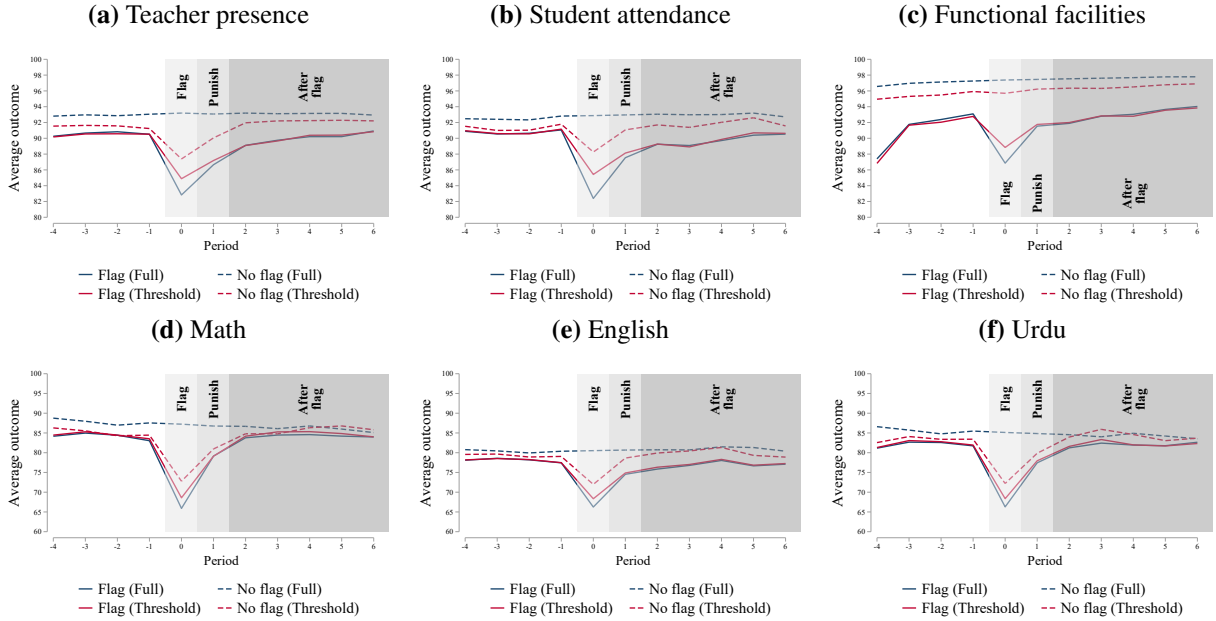
To explore this possibility, we use the school-by-month data to build a stacked dataset where, for each monthly markaz flagging event, we make a sub-dataset consisting of schools in a flagged markaz at the time of the event but not flagged before in some arbitrary number of pre-periods. For comparison, we consider those schools never in a flagged markaz during the sub-dataset time window of interest. We then stack all the sub-datasets together so the flagging event is centered in relative time, and we can compare the evolution of schools in flagged and unflagged maraakiz. For consistency with our analysis above, we can also identify those schools in maraakiz close to the flagging thresholds for each outcome. More precisely, we identify the maraakiz within an optimal bandwidth on either side of the flagging threshold in event-time 0 (Calonico, Cattaneo and Farrell, 2020), as in Section 4.[32]

Figure 6 presents the evolution of the stacked outcomes in relative time, anchored around periods of lag. Blue lines show the full stacked sample. Red lines plot the evolution for the threshold/optimal bandwidth sample around the flagging threshold. Solid lines represent schools in flagged maraakiz. Dotted lines are schools in non-flagged maraakiz. We highlight three periods corresponding to the month in which the data is collected and the flag is defined – *Flag*, the month in which these are reported to oversight committees and punishments occur – *Punish*, and the period after the flagging events, where we assess the impact of treatment – *After flag*.

[32] We obtain optimal bandwidths separately for each event panel to build a stacked-threshold sample. The average optimal bandwidth for teacher presence is 1.98 pp, for student attendance is 2.46 pp, for functional facilities is 4.18 pp, for Math is 3.83 pp, for English is 3.33 pp, and for Urdu is 3.48 pp.

Focusing first on the blue lines that compare all flagged maraakiz (solid) with non-flagged maraakiz (dotted), we observe that flagged and non-flagged marakiz 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 increased command-and-control accountability. This is similar to an Ashenfelter dip (Ashenfelter, 1978; Ashenfelter and Card, 1984; Heckman and Smith, 1999), where self-selection into the treatment happens because of a negative shock.

**Figure 6:** 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. 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. Scores variables are measured as the share of correct answers of students in standardized exams.

A senior manager following the trajectory of school outcomes in either the full or threshold samples of flagged maraakiz across the flagging and punishment periods and beyond would observe the trajectories illustrated by the solid lines in Figure 6. A naive interpretation of these dynamics is that the maraakiz are seemingly responsive to the flagging and punishment periods. However, by comparing the dynamics of the flagged maraakiz with those that were not flagged (the dotted lines), one can perceive little impact of flagging on the overall trajectories of outcomes. Flagged maraakiz do no better than non-flagged maraakiz, and typically revert to their mean level of performance before the period 0 shock by the 4th proceeding month. In addition, even non-flagged maraakiz exhibit a dip in performance in the flagging period and revert to historical trends, further showing

that flagging itself is not contributing to recovery.

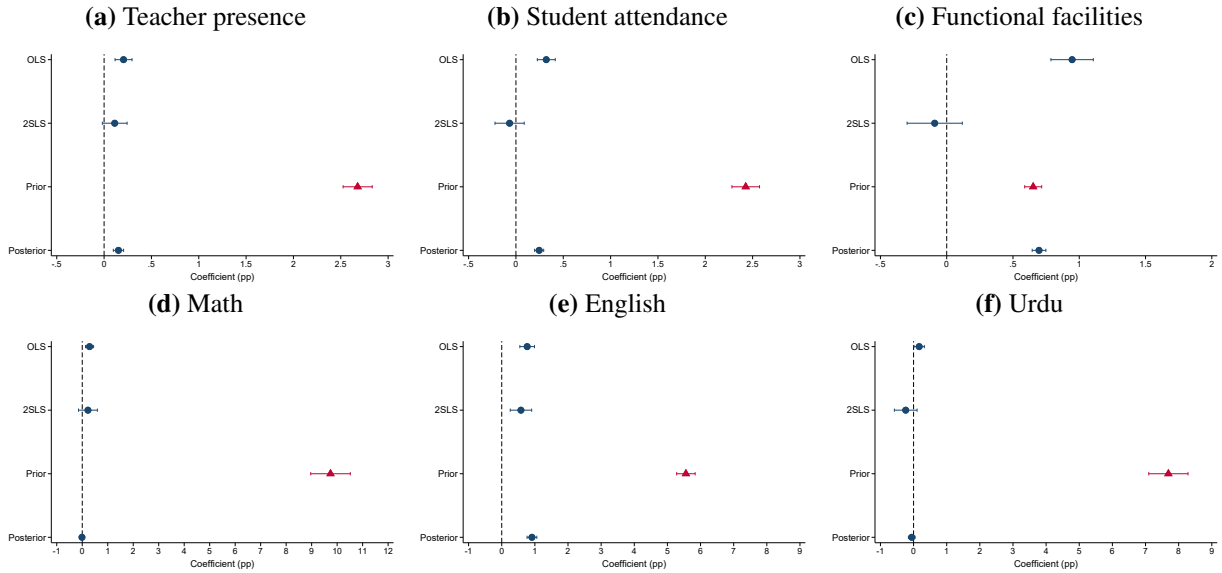
Appendix D presents the corresponding statistical assessment of these trends by applying a stacked (Cengiz et al., 2019; Baker, Larcker and Wang, 2022) difference-in-discontinuities approach (Grembi, Nannicini and Troiano, 2016) to the setup described here. As a natural extension, Appendix D also shows the results of estimating markaz-by-month effects using regression discontinuity design. Overall, the analysis implies the observed dynamics are equivalent to a reversion to the mean. By naively interpreting the natural reversion of school outcomes to their pre-shock means, senior managers may incorrectly associate better public sector outcomes with their own command-and-control interventions.

**Interpreting Null Effects of the Scheme in a Bayesian Framework** If senior managers misinterpret reversion to the mean as the positive effect of scheme, how should our data analysis update senior managers' prior beliefs? We use a Bayesian framework (Iacovone, McKenzie and Meager, 2025) that incorporates our estimates of naive treatment effects as the senior managers' beliefs. We estimate the reversion to the mean as the prior on the effects of the intensity of flagging using the specification described in Appendix D. The estimated coefficient and standard deviation are fed into the Bayesian analysis as normally distributed priors. Next, we estimate equation 1 assuming outcomes follow Gaussian likelihood.[33] Posterior distributions (effects updated by the priors) are calculated using Markov Chain Monte Carlo simulations.

Figure 7 compares frequentist (OLS and 2SLS) estimates with Bayesian priors and posteriors across six outcome variables. The priors reflect a strong initial belief that the scheme substantially improves outcomes. However, the posterior estimates, updated after incorporating the observed data, are consistently close to zero and align closely with the OLS and 2SLS results. This pattern indicates that the data provide little support for the prior assumption of large positive effects. In essence, the Bayesian updating process strongly downweights the optimistic prior beliefs, suggesting that the scheme had no statistically or substantively meaningful impact on the education system.

[33] We place a distribution on the rest of the parameters in our model. Fixed effects are nuisance parameters, and regularized toward zero by placing a hierarchical normal prior that shrinks them towards their mean.Inference is performed jointly over all parameters so fixed effects are not partialed out. In small samples, this might result in overfitting and noisy estimates (Athey and Imbens, 2017). Our sample size makes this problem unlikely, but the hierarchical regularization improves precision due to the high-dimensionality of our fixed effects. For the constant, we define a normal prior centered at the baseline outcome mean and scaled by its standard deviation.

**Figure 7:** Bayesian prior and posterior comparison




*Note:* This figure presents the coefficient point and 95% confidence interval from estimating different regressions. Coefficients for *OLS* and *2SLS* are estimated using equation 3 and reported in Table 4. Coefficient for senior managers *Prior* estimated using equation 4, estimating the effect of monthly flagging between the flag and punish periods as described in Appendix D. It captures the naive belief on the effectiveness of flagging. *Posterior* is the estimate for the effects using a Bayesian approach. We use a Gaussian model to define the likelihood on the outcomes. We regularize and shrink fixed effects towards their mean using a hierarchical normal prior with mean centered at zero and standard deviation scaled by the outcome standard deviation. For the constant, we define a normal prior centered at the baseline outcome mean and scaled by its standard deviation. Posterior distributions are calculated by Markov Chain Monte Carlo simulations using the Metropolis-Hastings algorithm.

# 7 Conclusions

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 these data through top-down methods of exerting control and accountability over 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 the main accountability instrument used in applying 'command-and-control' in government administration by evaluating a system from Punjab province in Pakistan that alerted senior government managers to poorly performing jurisdictions. Despite intensive flagging of poor performance leading to de facto accountability along the bureaucratic hierarchy, we detect 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 negligible impacts on any aspect of government functioning beyond a slight increase in government funding. Our data allow us to make these claims with a high degree of precision. Taken together, the results suggest that centralized command-and-control management approaches to accountability and control struggle to effectively manage unpredictable delivery environments.

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, a 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. We provide a contribution to that debate.

A natural question that arises from these findings is why a management approach with such limited effects persists as a phenomenon observed in public services around the world. By capitalizing on the fine-grained temporal nature of our data, we highlight that a naive evaluation of the scheme may lead senior managers to believe their interventions have subsequent positive impacts on school outcomes. In contrast, in all but one period and for one outcome, we do not observe a transition after flagging that differs from an organic reversion-to-the-mean.

In conclusion, our paper provides a detailed evaluation of a fundamental component of centralized accountability systems debated in the literature (Kane and Staiger, 2002; Besley and Coate, 2003; Bardhan, 2002; Dal Bó et al., 2021). Our results support the perspective that accountability approaches within 'command-and-control' systems fail to induce economically-meaningful changes throughout a public sector hierarchy.

## References


**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.

**Andrews, Isaiah, James H. Stock, and Liyang Sun.** 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics*, 11(Volume 11, 2019): 727–753.

**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.

**Athey, Susan, and Guido W Imbens.** 2017. "The econometrics of randomized experiments." In *Handbook of economic field experiments*. Vol. 1, 73–140. Elsevier.

**Azoulay, Pierre, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat.** 2019. "Public R&D investments and private-sector patenting: evidence from NIH funding rules." *The Review of economic studies*, 86(1): 117–152.

**Bahar, Dany, Prithwiraj Choudhury, Do Yoon Kim, and Wesley W Koo.** 2023. "Innovation on wings: Nonstop flights and firm innovation in the global context." *Management Science*, 69(10): 6202–6223.

**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.

**Baskaran, Thushyanthan, Zohal Hessami, and Sebastian Schirner.** 2024. "Young versus old politicians and public spending priorities." *Journal of Economic Behavior & Organization*, 225: 88–106.

**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.

**Bhalotra, Sonia, and Irma Clots-Figueras.** 2014. "Health and the political agency of women." *American Economic Journal: Economic Policy*, 6(2): 164–197.

**Bhalotra, Sonia, Irma Clots-Figueras, and Lakshmi Iyer.** 2021. "Religion and abortion: The role of politician identity." *Journal of Development Economics*, 153: 102746.

**Bhalotra, Sonia, Irma Clots-Figueras, Guilhem Cassan, and Lakshmi Iyer.** 2014. "Religion, politician identity and development outcomes: Evidence from India." *Journal of Economic Behavior & Organization*, 104: 4–17.

**Björkman, Martina, and Jakob Svensson.** 2009. "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda*." *The Quarterly Journal of Economics*, 124(2): 735–769.

**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.

**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.

**Borusyak, Kirill, and Matan Kolerman-Shemer.** 2024. "Regression discontinuity aggregation, with an application to the union effects on inequality." *arXiv preprint arXiv:2501.00428*.

**Borusyak, Kirill, and Peter Hull.** 2023. "Nonrandom exposure to exogenous shocks." *Econometrica*, 91(6): 2155–2185.

**Brierley, Sarah, Kenneth Lowande, Rachel Augustine Potter, and Guillermo Toral.** 2023. "Bureaucratic politics: Blind spots and opportunities in political science." *Annual Review of Political Science*, 26(1): 271–290.

**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.

**Campante, Filipe, and David Yanagizawa-Drott.** 2018. "Long-range growth: economic development in the global network of air links." *The Quarterly Journal of Economics*, 133(3): 1395–1458.

**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.

**Cattaneo, Matias D, and Rocio Titiunik.** 2022. "Regression discontinuity designs." *Annual Review of Economics*, 14: 821–851.

**Cattaneo, Matias D, Brigham R Frandsen, and Rocio Titiunik.** 2015. "Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate." *Journal of Causal Inference*, 3(1): 1–24.

**Cattaneo, Matias D, Michael Jansson, and Xinwei Ma.** 2020. "Simple local polynomial density estimators." *Journal of the American Statistical Association*, 115(531): 1449–1455.

**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.

**Chen, Yvonne Jie, Pei Li, and Yi Lu.** 2018. "Career concerns and multitasking local bureaucrats: Evidence of a target-based performance evaluation system in China." *Journal of Development Economics*, 133: 84–101.

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

**Clots-Figueras, Irma.** 2011. "Women in politics: Evidence from the Indian States." *Journal of public Economics*, 95(7-8): 664–690.

**Clots-Figueras, Irma.** 2012. "Are Female Leaders Good for Education? Evidence from India." *American Economic Journal: Applied Economics*, 4(1): 212–44.

**Craig, Steven G., Scott A. Imberman, and Adam Perdue.** 2015. "Do administrators respond to their accountability ratings? The response of school budgets to accountability grades." *Economics of Education Review*, 49: 55–68.

**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.

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

**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.

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

**Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, Gareth Nellis, Claire L. Adida, Eric Arias, Clara Bicalho, Taylor C. Boas, Mark T. Buntaine, Simon Chauchard, Anirvan Chowdhury, Jessica Gottlieb, F. Daniel Hidalgo, Marcus Holmlund, Ryan Jablonski, Eric Kramon, Horacio Larreguy, Malte Lierl, John Marshall, Gwyneth McClendon, Marcus A. Melo, Daniel L. Nielson, Paula M. Pickering, Melina R. Platas, Pablo Querubín, Pia Raffler, and Neelanjan Sircar.** 2019. "Voter Information Campaigns and Political Accountability: Cumulative Findings from a Preregistered Meta-Analysis of Coordinated Trials." *Science Advances*, 5(7): eaaw2612.

**Easterly, William.** 2008. "Institutions: Top Down or Bottom Up?" *American Economic Review*, 98(2): 95–99.

**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.

**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.

**Folke, Olle.** 2014. "Shades of brown and green: party effects in proportional election systems." *Journal of the European Economic Association*, 12(5): 1361–1395.

**Freier, Ronny, and Christian Odendahl.** 2015. "Do parties matter? Estimating the effect of political power in multi-party systems." *European Economic Review*, 80: 310–328.

**Geys, Benny, Zuzana Murdoch, and Rune J Sørensen.** 2024. "Public Employees as Elected Politicians: Assessing Direct and Indirect Substantive Effects of Passive Representation." *The Journal of Politics*, 86(1): 170–182.

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

**Grembi, Veronica, Tommaso Nannicini, and Ugo Troiano.** 2016. "Do fiscal rules matter?" *American Economic Journal: Applied Economics*, 1–30.

**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.

**Hyytinen, Ari, Jaakko Meriläinen, Tuukka Saarimaa, Otto Toivanen, and Janne Tukiainen.** 2018. "Public employees as politicians: Evidence from close elections." *American Political Science Review*, 112(1): 68–81.

**Iacovone, Leonardo, David McKenzie, and Rachael Meager.** 2025. "Bayesian impact evaluation with informative priors: An application to a colombian management and export improvement program." *Econometrica*, 93(5): 1915–1935.

**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.** 2025. "Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan." *American Economic Review*.

**Kleibergen, Frank, and Richard Paap.** 2006. "Generalized reduced rank tests using the singular value decomposition." *Journal of econometrics*, 133(1): 97–126.

**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.** 2023. "Improving service delivery via top-down data-driven accountability: Reform enactment of the Education Road Map in Pakistan." DeliverEd Initiative 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.

**Meriläinen, Jaakko.** 2022. "Political selection and economic policy." *The Economic Journal*, 132(648): 3020–3046.

**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.

**Nellis, Gareth, and Niloufer Siddiqui.** 2018. "Secular party rule and religious violence in Pakistan." *American Political Science Review*, 112(1): 49–67.

**Nellis, Gareth, Michael Weaver, Steven C Rosenzweig, et al.** 2016. "Do parties matter for ethnic violence? Evidence from India." *Quarterly Journal of Political Science*, 11(3): 249–277.

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

**Priyanka, Sadia.** 2020. "Do female politicians matter for female labor market outcomes? Evidence from state legislative elections in India." *Labour Economics*, 64: 101822.

**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.

**Sørensen, Rune J.** 2023. "Educated politicians and government efficiency: Evidence from Norwegian local government." *Journal of Economic Behavior & Organization*, 210: 163–179.

**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

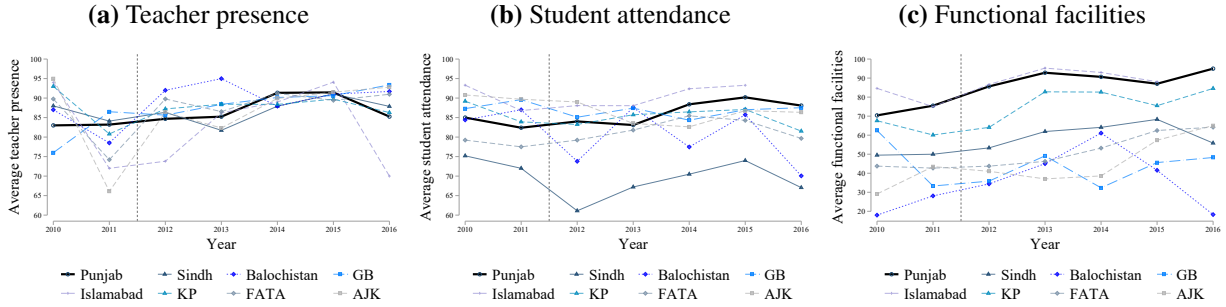
## A Data and Design Details

### A.1 Broad context of education outcomes in Pakistan

As argued in the main paper, the command-and-control intervention was a major initiative of the Punjab Government. Similar schemes were not implemented in other provinces of Pakistan during the same period of time. As such, a broad reflection on the scheme can be had by comparing the trajectory of education in Punjab to that in other provinces. We recover province-level data for the period 2010-2016 from the Annual Status of Education Report - ASER - Pakistan (aserpakistan.org), which have conducted independently and consistently household and school surveys to assess education progress in the country.

Figure A1 shows the average trends in educational outcomes in all Pakistan provinces. Note that most provinces are improving or trending in a similar way to Punjab (darker blue line). So despite some underperforming provinces, most of the country faces similar evolving trends.

**Figure A1:** Pakistan provinces average outcomes trends




*Note:* The figure shows the average trends of education outcomes in all Pakistan provinces using data from ASER Pakistan (aserpakistan.org), for the period 2010-2016, which have been independently and consistently conducting household and school surveys to assess the education advancements in the country. 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. 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.

### A.2 Color-coded performance thresholds

Teacher presence was coded red when it fell below 86%, orange when it was between 86% and 90%, and green when it was 90% or higher. Functional facilities thresholds were 90% and 95%. For both variables, the thresholds were the same across all administrative units and time periods.

The thresholds for student attendance varied between districts and months. The districts were divided into three categories, A, B, and C, where category A consisted of historically the highest performing districts, category C consisted of historically lowest performing districts, and category B consisted of the rest. Furthermore, the months in the year were divided into high attendance (December-March), and low attendance (April-November). Different thresholds were established for each category of districts and group of months. For category A districts during December-March, student attendance was coded red if it was below 89%, orange if it was between 89% and 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.

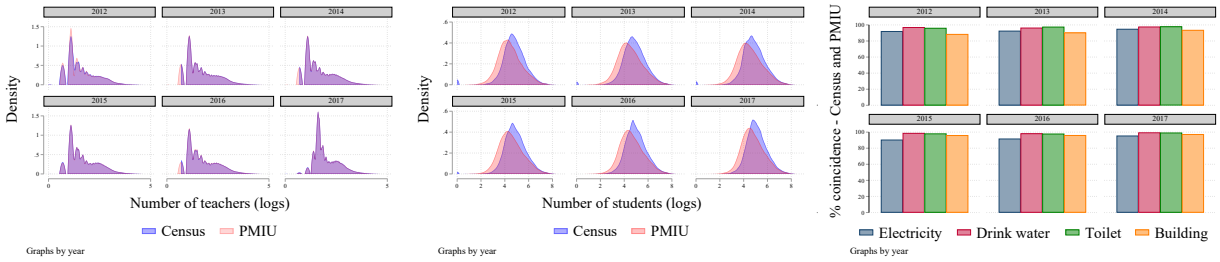
### A.3 Data compliance

We have data-pack reports for 60 months from December 2011 to May 2018, which represent 100% of the reporting.[34] We compare these datapack reports with census data to assess their quality. The Annual School Census is used to collect comprehensive statistics on the education sector in Punjab. It is the government's primary source of information for public policy and resource allocation. Because it is collected yearly, there is a longer training period for data collectors and a longer span for data validation and correction. As such, data collected by the census is high quality and passes multiple validation checks. We assess the quality of the data collected in the data-packs by comparing it against the census in the month where the census information was collected (October). However, since data are not collected on the same day, nor by the same sources, there can be some measurement error. Figure A2 compares the distribution of the variables reported by both sources. Panels (a) show for teacher presence that the data-pack and census data overlap almost completely, suggesting that the information collected in the data-packs mapped consistently the population behaviour reported by the census. Panel (b) also shows an almost full overlap in student attendance across the two information sources. Panel (c) plots the percentage of schools where the functional infrastructure coincides, which is near 100% for all the indicators. As such, there is no systematic manipulation of the monthly performance measures, further supporting the reliability of the monthly data for the analysis.

[34] June, July, and August are not included as schools are not in session

**Figure A2:** Data validation - monthly PMIU vs. Census

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




*Note:* This figure compares October PMIU data and corresponding school-level quantities from the Annual School Census. 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. 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).

## A.4 Descriptive statistics

**Table A1:** Descriptive statistics by flagging status

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Panel A: Markaz-level variables** | | | | | | | | |
| | Mean | Median | Std. Dev. | Obs | Mean | Median | Std. Dev. | Obs |
| Number of schools | 22 | 16 | 21 | 142,962 | 22 | 16 | 21 | 142,962 |
| Proportion elementary | 80 | 100 | 40 | 142,962 | 80 | 100 | 40 | 142,962 |
| Outcomes (0-100) | | | **No flag** | | | | **Flag** | |
| Teacher presence | 93 | 94 | 4.4 | 104,667 | 80 | 83 | 8.4 | 10,031 |
| Student attendance | 91 | 92 | 6 | 99,045 | 80 | 82 | 7.9 | 15,625 |
| Functional facilities | 95 | 98 | 11 | 98,953 | 81 | 84 | 11 | 15,693 |
| Math score | 87 | 88 | 6.4 | 66,713 | 64 | 66 | 5.5 | 2,392 |
| English score | 80 | 80 | 6.4 | 54,364 | 64 | 66 | 5.4 | 14,741 |
| Urdu score | 85 | 86 | 6.4 | 66,011 | 65 | 67 | 5.8 | 3,094 |
| **Panel B: School-level variables** | | | | | | | | |
| | Mean | Median | Std. Dev. | Obs | Mean | Median | Std. Dev. | Obs |
| Number of teachers | 4.5 | 3 | 3.8 | 2,627,487 | 4.5 | 3 | 3.8 | 2,627,487 |
| Number of students | 109 | 78 | 102 | 2,632,372 | 109 | 78 | 102 | 2,632,372 |
| Outcomes (0-100) | | | **No flag** | | | | **Flag** | |
| Teacher presence | 92 | 100 | 15 | 2,378,448 | 84 | 100 | 22 | 244,600 |
| Student attendance | 89 | 93 | 12 | 2,175,245 | 81 | 85 | 17 | 451,303 |
| Functional facilities | 92 | 100 | 18 | 2,134,405 | 83 | 100 | 23 | 448,187 |
| Math score | 87 | 92 | 14 | 905,036 | 68 | 67 | 21 | 24,760 |
| English score | 79 | 83 | 18 | 725,631 | 66 | 67 | 20 | 204,147 |
| Urdu score | 85 | 89 | 15 | 890,076 | 68 | 71 | 20 | 39,694 |
| **Panel C: District-level variables** | | | | | | | | |
| | Mean | Median | Std. Dev. | Obs | Mean | Median | Std. Dev. | Obs |
| Outcomes (0-100) | | | **Top 5** | | | | **Bottom 5** | |
| Overall score | 94 | 95 | 3.8 | 70 | 78 | 78 | 10 | 70 |
| New position | 7.7 | 0 | 27 | 504 | 8.3 | 0 | 28 | 504 |

*Notes:* The unit for outcomes in Panel A is outcome-markaz-month; in Panel B it is outcome-school-month. Outcomes are measured in percentages from 0 to 100. 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 datapack 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.

# B Intensity of Exposure to Flagging

**Table B1:** Descriptive statistics - flagging in the threshold

| **Schooling outcome - flagging variable** | Mean | Std. Dev. | Min. | Max. | Obs. |
|---|---|---|---|---|---|
| Teacher presence - # times close to the threshold | 0.613 | 1.161 | 0 | 8 | 16,126 |
| Teacher presence - # times flagged close to the threshold | 0.214 | 0.592 | 0 | 8 | 16,126 |
| Student attendance - # times close to the threshold | 0.936 | 1.524 | 0 | 8 | 16,129 |
| Student attendance - # times flagged close to the threshold | 0.358 | 0.825 | 0 | 8 | 16,129 |
| Functional facilities - # times close to the threshold | 0.718 | 1.463 | 0 | 8 | 16,114 |
| Functional facilities - # times flagged close to the threshold | 0.239 | 0.773 | 0 | 8 | 16,114 |
| Match score - # times close to the threshold | 0.171 | 0.468 | 0 | 8 | 9,482 |
| Match score - # times flagged close to the threshold | 0.055 | 0.260 | 0 | 8 | 9,482 |
| English score - # times close to the threshold | 1.193 | 1.427 | 0 | 8 | 9,482 |
| English score - # times flagged close to the threshold | 0.472 | 0.794 | 0 | 8 | 9,482 |
| Urdu score - # times close to the threshold | 0.302 | 0.630 | 0 | 7 | 9,482 |
| Urdu score - # times flagged close to the threshold | 0.086 | 0.304 | 0 | 3 | 9,482 |

*Notes:* The unit for the variables is markaz-year. Number of times close to the threshold refers to the number of months in the school-year that the markaz performance was around the optimal bandwidth around the flagging threshold. Number of times flagged close to the threshold refers then to cases where the markaz was flagged with a performance just below the flagging threshold. The outcome at the start of each row refers to the respective variable in which flagging and times in the threshold is being measured.

**Table B2:** Intensity of exposure to flagging - balance test maraakiz sample

**Panel A: School outcomes**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ (z-score) | 0.163*** | 0.036 | -0.017 | -0.191** | -0.163** | -0.173* |
| | (0.048) | (0.073) | (0.055) | (0.088) | (0.075) | (0.101) |
| N. of obs. | 204,224 | 204,224 | 204,398 | 204,398 | 201,600 | 201,600 |
| Number markaz | 2,871 | 2,871 | 2,871 | 2,871 | 2,871 | 2,871 |
| Mean Dep. Var. | 91.6 | 91.6 | 87.6 | 87.6 | 90.2 | 90.2 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 |
| First stage F-stat | | 187.0 | | 222.3 | | 319.9 |

**Panel B: Student scores**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Math$_{t+1}$ | | English$_{t+1}$ | | Urdu$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ (z-score) | 0.269*** | 0.183 | 0.764*** | 0.579*** | 0.176** | -0.226 |
| | (0.082) | (0.190) | (0.113) | (0.165) | (0.081) | (0.175) |
| N. of obs. | 57,656 | 57,656 | 57,654 | 57,654 | 57,655 | 57,655 |
| Number markaz | 1,837 | 1,837 | 1,837 | 1,837 | 1,837 | 1,837 |
| Mean Dep. Var. | 86.9 | 86.9 | 76.5 | 76.5 | 84.6 | 84.6 |
| Mean # Times flagged | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 35.5 | | 156.7 | | 53.3 |
| 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 |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. This table present the results from estimating equation 3 on the sample used for testing orthogonality of the instrument in Table 3, where we use only 80% of maraakiz. This table is then a robustness test for the results presented in Table 4. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A reports on schooling outcomes. Panel B reports on student scores. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

**Table B3:** Intensity of exposure to flagging - RDA estimator

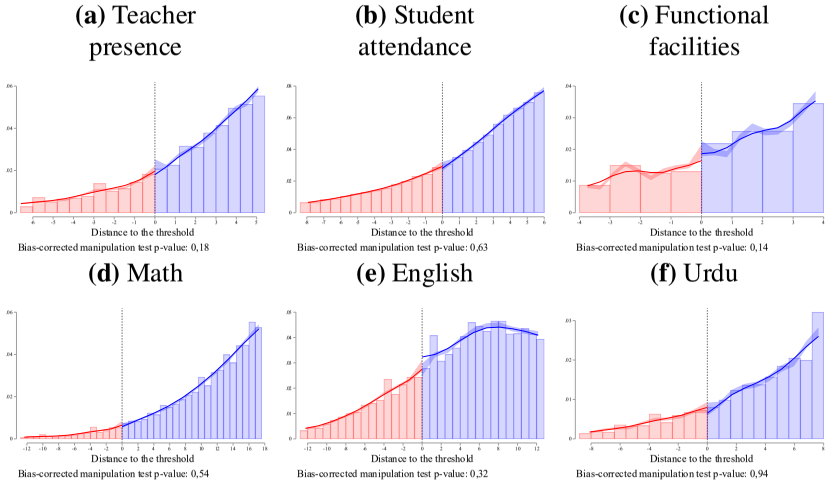
| | Dependent variables (range: 0-100) | | |
|---|---|---|---|
| | Teacher presence$_{t+1}$ | Student attendance$_{t+1}$ | Functional facilities$_{t+1}$ |
| | 2SLS | 2SLS | 2SLS |
| | (1) | (2) | (3) |
| # Times flagged$_t$ (z-score) | 0.102 | -0.002 | -0.014 |
| | (0.067) | (0.073) | (0.107) |
| N. of obs. | 257,592 | 257,865 | 254,058 |
| Number markaz | 3,567 | 3,567 | 3,567 |
| Mean Dep. Var. | 91.6 | 87.6 | 90.2 |
| First stage F-stat | 1762.1 | 2147.1 | 2353.6 |
| Markaz FE | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes |
| District time trends | Yes | Yes | Yes |
| RDA Controls | Yes | Yes | Yes |

**Panel B: Student scores**

| | Dependent variables (range: 0-100) | | |
|---|---|---|---|
| | Math$_{t+1}$ | English$_{t+1}$ | Urdu$_{t+1}$ |
| | 2SLS | 2SLS | 2SLS |
| | (1) | (2) | (3) |
| # Times flagged$_t$ (z-score) | 0.113 | 0.498*** | -0.628** |
| | (0.299) | (0.156) | (0.285) |
| N. of obs. | 67,385 | 67,383 | 67,384 |
| Number markaz | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. | 86.9 | 76.5 | 84.6 |
| First stage F-stat | 48.9 | 1342.6 | 77.6 |
| Markaz FE | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes |
| District time trends | Yes | Yes | Yes |
| RDA Controls | Yes | Yes | Yes |

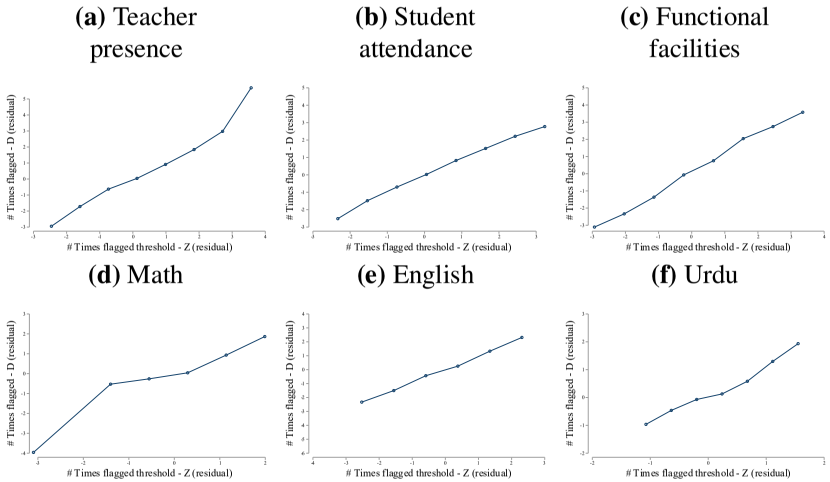
*Notes:* The unit of analysis is the school-year. This table presents the results from using the RDA –Regression Discontinuity Aggregation– method proposed by Borusyak and Kolerman-Shemer (2024), which intends to simulate the local linear estimation of RD designs. We estimate equation 3 using as instrument the share of months that a markaz was flagged while being close to the threshold, and controlling by the share of months in a year a markaz was close to the threshold instead of the number of months close to the threshold ($\theta_{m,d,t}$), to be consistent with the RDA. In addition, and following the RDA, we control for i) the weighted average markaz performance for months where performance was close to the threshold, and ii) the weighted average markaz performance for months where markaz was flagged while being close to the threshold. Performance in i) and ii) are recentered by substracting the threshold value. The weights are the same for each month of the school year and sum to one. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Results not reported for student scores as controls overlap between them. Times flagged$_t$ (z-score) is the share of months a markaz was flagged in year $t$ in the outcome reported at the top of the column. The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

**Figure B1:** Manipulation around the flagging threshold test



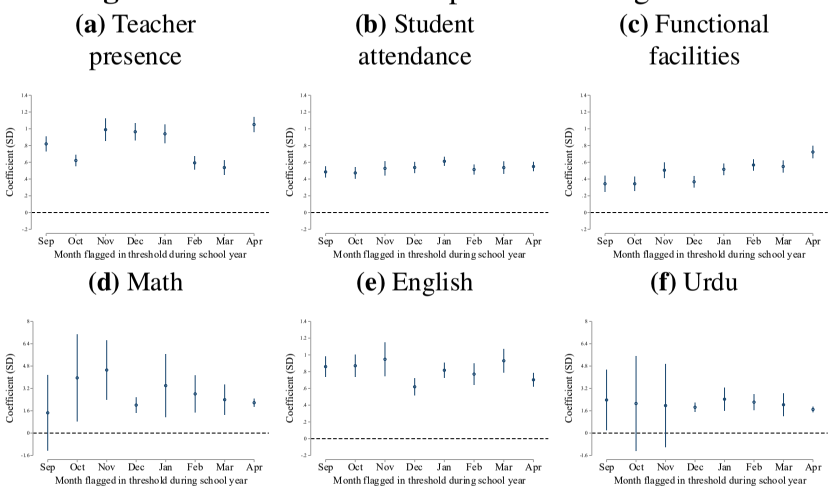

*Note:* This figure reports the manipulation test proposed by Cattaneo, Jansson and Ma (2020) using the command *rddensity* in stata for each outcome, using a unrestricted density estimation, triangular kernel, and jackknife standard errors. The p-value corresponds to the bias-corrected p-value from estimating the discontinuity in the outcome density around the flagging threshold, reported at the bottom of the figure. The test is performed using markaz-by-month data, which is the relevant level of flagging. For the functional facilities outcome, there are specific peaks in the distribution due to functional facilities not being clearly defined as a continuous variable. At the school level, it is measured as the share of functional facilities between four possible options. We report larger bins and the results from a restricted test to account for the distribution of the variable.

**Figure B2:** Monotonicity assumption



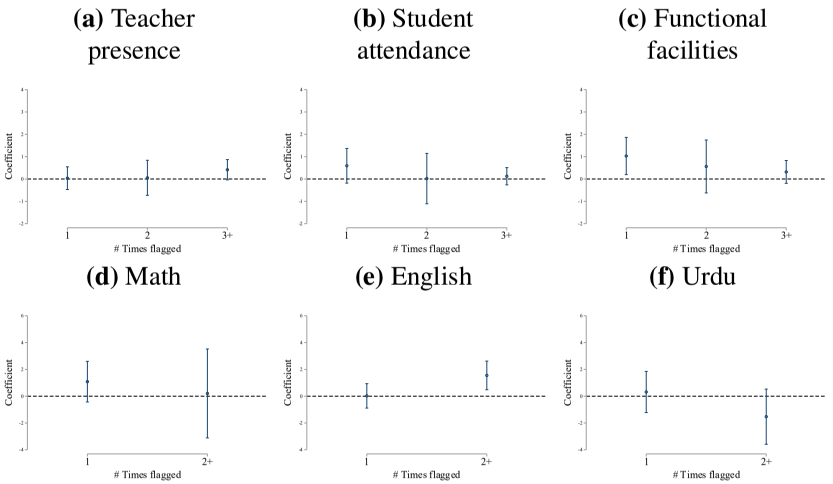

*Note:* This figure reports the correlation for the # Times flagged and the # Times flagged threshold after residualizing the fixed effects defined in equation 2 to test for the monotonicity assumption of instrument. Under monotonicity, each additional flagging close to the threshold must lead to no less than an extra flag in general, so a positive relationship exists between both variables. The correlations are estimated in the markaz-by-school year data. The points are grouped in equal-length bins. Each panel reports on the respective outcome flagging.

**Figure B3:** Relevance assumption - first stage results




*Note:* This figure presents results from the $\gamma_j$ coefficients of estimating equation 2, where $S_j$ is for individual dummies for each month $j$ of the school year in which a markaz was flagged by being close to the threshold. The coefficients test for the relevance assumption of the first stage for results in Table 4. The outcome is measured as the (normalized) # times a markaz was flagged in a year. Thus, the coefficients are interpreted using standard deviations (SD). May is used as the base month, so no coefficient has been reported for it. Each panel reports on the respective outcome flagging. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students in each school. Error bars at the 95 percent level, clustered at the markaz level, are presented for each coefficient.

**Figure B4:** Intensity of exposure to flagging - non-linearity and alternative endogenous variable definition




*Note:* This figure presents the results from estimating equation 3 under different definitions of the endogenous and instrumental variables to capture non-linearities in the effects, as a robustness test for the results reported in Table 4. Panels show the point estimates defining the endogenous variable equal to one if the markaz was flagged once, equal to two if the markaz was flagged twice, and equal to three if the markaz was flagged thrice or more. For student scores variables, we define the endogenous variable as one if the markaz was flagged once, and equal to two if the markaz was flagged twice or more. The instrument for the first stage in each case is defined in the same way but for the # Times flagged close to the threshold. The base category for all definitions is equal to zero if there is no flagging. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students in each school. Error bars at the 95 percent level, clustered at the markaz level, are presented for each coefficient.

**Table B4:** Intensity of exposure to flagging - alternative instrument definition

**Panel A: School outcomes**

| | Dependent variables (range: 0-100) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | | Student attendance$_{t+1}$ | | | Functional facilities$_{t+1}$ | | |
| | Linear | Ever flag | Median | Linear | Ever flag | Median | Linear | Ever flag | Median |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| # Times flagged$_t$ (z-score) | 0.124* | 0.110 | 0.130* | -0.010 | 0.122 | 0.003 | -0.036 | 0.267** | 0.111 |
| | (0.066) | (0.080) | (0.070) | (0.076) | (0.100) | (0.085) | (0.108) | (0.135) | (0.117) |
| N. of obs. | 257,592 | 257,592 | 257,592 | 257,865 | 257,865 | 257,865 | 254,058 | 254,058 | 254,058 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 |
| Mean Dep. Var. | 91.6 | 91.6 | 91.6 | 87.6 | 87.6 | 87.6 | 90.2 | 90.2 | 90.2 |
| Mean # Times flagged | 0.87 | 0.87 | 0.87 | 1.59 | 1.59 | 1.59 | 1.61 | 1.61 | 1.61 |
| SD # Times flagged | 1.42 | 1.42 | 1.42 | 2.19 | 2.19 | 2.19 | 2.81 | 2.81 | 2.81 |
| First stage F-stat | 1784.2 | 1064.7 | 802.1 | 2162.6 | 1083.5 | 904.0 | 2421.0 | 1356.0 | 876.2 |

**Panel B: Student scores**

| | Dependent variables (range: 0-100) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Math$_{t+1}$ | | | English$_{t+1}$ | | | Urdu$_{t+1}$ | | |
| | Linear | Ever flag | Median | Linear | Ever flag | Median | Linear | Ever flag | Median |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| # Times flagged$_t$ (z-score) | 0.281 | 0.290 | 0.244 | 0.448*** | 0.361* | 0.481*** | -0.237 | -0.162 | -0.203 |
| | (0.190) | (0.211) | (0.195) | (0.154) | (0.218) | (0.165) | (0.171) | (0.178) | (0.169) |
| N. of obs. | 67,385 | 67,385 | 67,385 | 67,383 | 67,383 | 67,383 | 67,384 | 67,384 | 67,384 |
| Number markaz | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. | 86.9 | 86.9 | 86.9 | 76.5 | 76.5 | 76.5 | 84.6 | 84.6 | 84.6 |
| Mean # Times flagged | 0.038 | 0.038 | 0.038 | 0.34 | 0.34 | 0.34 | 0.065 | 0.065 | 0.065 |
| SD # Times flagged | 0.24 | 0.24 | 0.24 | 0.96 | 0.96 | 0.96 | 0.33 | 0.33 | 0.33 |
| First stage F-stat | 220.4 | 213.8 | 112.8 | 1382.8 | 516.0 | 540.9 | 363.5 | 366.4 | 191.6 |
| Markaz FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| District time trends | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. This table presents the results from estimating equation 3 through 2SLS, varying the definition of the instrumental variable as a robustness test for Table 4. The *Main* instrument is defined as dummy variables for each month of a year equal to one if in such month a markaz was flagged while close to the threshold. This table report the following variations for the instrument – *Linear*: normalized number of times flagged close to the threshold sample. *Ever*: a dummy equal to one if the markaz reported to be flagged at least once in a year, zero otherwise. *Median*: a categorical variable dividing between markaz flagged above/below the median of the distribution. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A reports on schooling outcomes. Panel B reports on student scores. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

**Table B5:** Intensity of exposure to flagging - other performance metrics

**Panel A: Best performance month**

| | Dependent variables (range: 0-100) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | | Math$_{t+1}$ | | English$_{t+1}$ | | Urdu$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.11*** | 0.094** | 0.24*** | 0.0059 | 0.82*** | 0.13* | -0.13 | 0.39* | 0.11 | 0.047 | -0.15* | -0.12 |
| | (0.029) | (0.044) | (0.043) | (0.066) | (0.068) | (0.073) | (0.096) | (0.21) | (0.14) | (0.20) | (0.090) | (0.21) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 | 67,385 | 67,385 | 67,383 | 67,383 | 67,384 | 67,384 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. | 98.8 | 98.8 | 94.9 | 94.9 | 93.3 | 93.3 | 96.6 | 96.6 | 91.4 | 91.4 | 95.2 | 95.2 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 210.8 | | 258.6 | | 382.5 | | 34.8 | | 157.3 | | 53.3 |

**Panel B: Worst performance month**

| | Dependent variables (range: 0-100) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | | Math$_{t+1}$ | | English$_{t+1}$ | | Urdu$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.395*** | 0.190 | 0.537*** | -0.144 | 0.610*** | -0.726 | 1.017*** | -0.055 | 1.445*** | 1.004*** | 0.740*** | -0.133 |
| | (0.107) | (0.173) | (0.124) | (0.191) | (0.216) | (0.509) | (0.128) | (0.289) | (0.170) | (0.263) | (0.128) | (0.253) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 | 67,385 | 67,385 | 67,383 | 67,383 | 67,384 | 67,384 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. | 75.7 | 75.7 | 75.7 | 75.7 | 85.2 | 85.2 | 71.8 | 71.8 | 58.3 | 58.3 | 69.5 | 69.5 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 210.8 | | 258.6 | | 382.5 | | 34.8 | | 157.3 | | 53.3 |
| Markaz FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| District time trends | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. This table present the results from estimating equation 3 on measures of performance different to the average yearly performance as used in Table 4. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A report the results for the outcomes defined as the best monthly performance in a school year. Panel B reports for the worst monthly performance in a school year. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

**Table B6:** Intensity of exposure to flagging by school performance

**Panel A: Best performance schools**

| | Dependent variables (range: 0-100) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | | Math$_{t+1}$ | | English$_{t+1}$ | | Urdu$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | -0.039*** | -0.026 | -0.129*** | -0.068** | 0.767*** | -0.205 | 0.015 | -0.120 | 0.199*** | 0.135 | 0.019 | -0.112 |
| | (0.013) | (0.021) | (0.019) | (0.027) | (0.108) | (0.143) | (0.044) | (0.093) | (0.077) | (0.116) | (0.041) | (0.078) |
| N. of obs. | 141,203 | 141,203 | 129,818 | 129,818 | 138,484 | 138,484 | 31,905 | 31,905 | 33,674 | 33,674 | 32,099 | 32,099 |
| Number markaz | 3,508 | 3,508 | 3,426 | 3,426 | 3,506 | 3,506 | 2,589 | 2,589 | 2,544 | 2,544 | 2,570 | 2,570 |
| Mean Dep. Var. | 91.6 | 91.6 | 87.6 | 87.6 | 90.2 | 90.2 | 86.9 | 86.9 | 76.5 | 76.5 | 84.6 | 84.6 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 240.9 | | 241.9 | | 328.2 | | 60.8 | | 166.2 | | 65.8 |

**Panel B: Worst performing schools**

| | Dependent variables (range: 0-100) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | | Math$_{t+1}$ | | English$_{t+1}$ | | Urdu$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.344*** | 0.208** | 0.602*** | 0.056 | 1.071*** | 0.034 | 0.189** | 0.306 | 0.439*** | 0.246 | 0.203** | -0.145 |
| | (0.062) | (0.087) | (0.054) | (0.083) | (0.100) | (0.123) | (0.087) | (0.226) | (0.116) | (0.164) | (0.084) | (0.173) |
| N. of obs. | 116,251 | 116,251 | 127,805 | 127,805 | 115,435 | 115,435 | 35,244 | 35,244 | 33,488 | 33,488 | 35,056 | 35,056 |
| Number markaz | 3,410 | 3,410 | 3,293 | 3,293 | 3,410 | 3,410 | 2,505 | 2,505 | 2,567 | 2,567 | 2,539 | 2,539 |
| Mean Dep. Var. | 91.6 | 91.6 | 87.6 | 87.6 | 90.2 | 90.2 | 86.9 | 86.9 | 76.5 | 76.5 | 84.6 | 84.6 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 166.3 | | 211.1 | | 360.6 | | 26.3 | | 131.5 | | 45.1 |
| Markaz FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| District time trends | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. This table show the results from estimating equation 3 on samples of schools to test how much the conclusions from Table 4 change under specific school characteristics. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A report for the schools performing above the median of the markaz average performance. Panel B reports for schools performing below the median of the markaz average performance. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

**Table B7:** Impacts of exposure to flagging - orange flagging

**Panel A: School outcomes**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ (z-score) | 0.151*** | 0.086 | 0.103** | 0.182*** | 0.207*** | 0.118 |
| | (0.037) | (0.054) | (0.047) | (0.066) | (0.048) | (0.076) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 |
| Mean Dep. Var. | 91.6 | 91.6 | 87.6 | 87.6 | 90.2 | 90.2 |
| Mean # Times flagged | 1.41 | 1.41 | 1.64 | 1.64 | 1.47 | 1.47 |
| SD # Times flagged | 1.62 | 1.62 | 1.73 | 1.73 | 2.36 | 2.36 |
| First stage F-stat | | 361.6 | | 467.8 | | 223.3 |

**Panel B: Student scores**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Math$_{t+1}$ | | English$_{t+1}$ | | Urdu$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ (z-score) | 0.298*** | 0.337** | 0.431*** | 1.446*** | 0.369*** | 0.115 |
| | (0.084) | (0.150) | (0.133) | (0.300) | (0.099) | (0.172) |
| N. of obs. | 67,385 | 67,385 | 67,383 | 67,383 | 67,384 | 67,384 |
| Number markaz | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. | 86.9 | 86.9 | 76.5 | 76.5 | 84.6 | 84.6 |
| Mean # Times flagged | 0.21 | 0.21 | 0.77 | 0.77 | 0.33 | 0.33 |
| SD # Times flagged | 0.66 | 0.66 | 1.63 | 1.63 | 0.90 | 0.90 |
| First stage F-stat | | 106.0 | | 54.0 | | 105.8 |
| 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 |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. Results from estimating equation 3. Flagging is defined as being below the green flagging threshold. That is, being flagged as 'orange', instead of 'red' as in the main analysis. Thus, not flagged units are coded as 'green' –best-performers– in the data-packs. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A reports on schooling outcomes. Panel B reports on student scores. # Times flagged$_t$ is the number of times a markaz was flagged 'orange' in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged 'orange' while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Scores are measured since 2016, and consist on the share of correct answers in standardized exams performed on a random sample of primary students. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

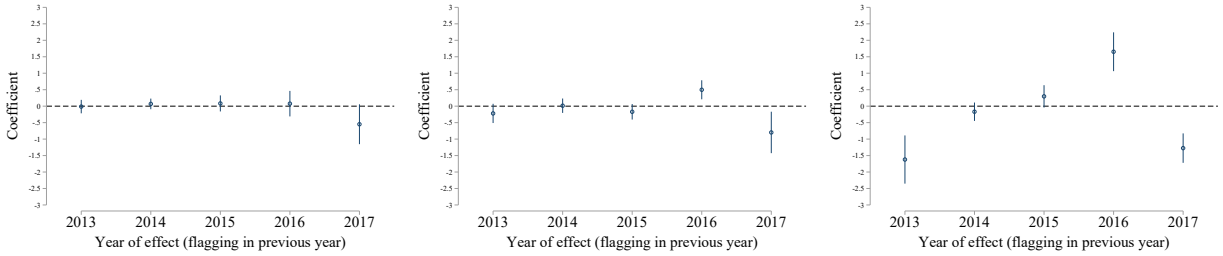
**Table B8:** Intensity of exposure to flagging - heterogeneity by district ranking

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ (z-score) | -0.046 | -0.210 | 0.815*** | 0.256 | 1.134*** | 0.635*** |
| | (0.121) | (0.188) | (0.127) | (0.246) | (0.230) | (0.236) |
| Bottom district$_t$ × # Times flagged$_t$ (z-score) | 0.325** | 0.424** | -0.497*** | -0.416 | -0.001 | -0.710*** |
| | (0.127) | (0.195) | (0.136) | (0.255) | (0.253) | (0.270) |
| Top district$_t$ × # Times flagged$_t$ (z-score) | 0.190 | 0.276 | -0.573*** | -0.282 | -0.381 | -0.815*** |
| | (0.128) | (0.199) | (0.135) | (0.256) | (0.249) | (0.255) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 |
| Mean Dep. Var. | 91.6 | 91.6 | 87.6 | 87.6 | 90.2 | 90.2 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 |
| 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 |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. This table show the results from estimating equation 3 adding an additional interaction for the district ranking to test for heterogeneity of the main results of Table 4. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Bottom (Top) district equals one if the average position of the district is in the bottom (top) five, relative to the districts outside the bottom and top positions. Year is measured as school-year (September to May). *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Standard errors clustered by markaz in parentheses. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

**Figure B5:** Impacts of exposure to flagging - effect by year

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




*Note:* This figure presents the results from estimating equation 3 by separating the endogenous variable and instrument by year, and using as base period 2021, the first year of the command-and-control system implementation. The figure test robustness test for the results reported in Table 4 from time effects. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. We do not report report results on scores as they are measured since 2016, thus we would only be able to estimate the coefficient for 2017. Also, we do not report the effect for 2018 as the data if incomplete for the academic year. Error bars at the 95 percent level, clustered at the markaz level, are presented for each coefficient. Coefficient interpreted as percentage points change from a one standrad deviation increase in flagging.

**Table B9:** Impacts of exposure to flagging - district ranking

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | |
| | 2SLS | 2SLS | 2SLS | 2SLS | 2SLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| District in the bottom 5 | -0.032 | | 0.582 | | 1.908 | |
| | (0.584) | | (0.666) | | (1.409) | |
| District in the top 5 | | 0.888** | | -0.276 | | 0.695 |
| | | (0.358) | | (1.028) | | (0.886) |
| N. of obs. | 344 | 344 | 346 | 346 | 353 | 353 |
| Number districts | 36 | 36 | 36 | 36 | 36 | 36 |
| Mean Dep. Var. | 90.8 | 90.8 | 87.6 | 87.6 | 91.4 | 91.4 |
| First stage F-stat | 133.4 | 372.5 | 113.1 | 189.3 | 177.2 | 322.9 |
| District FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes |
| District in the ranking threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the district-quarter. Estimates for the period between 2012 and 2015, for which we have data on district rankings. Because data on scores started to be available since 2016, we only report results on teacher presence, student attendance, and functional facilities. Results from estimating a modified version of equation 3: District in the top (bottom)$_t$ is a dummy equal to one if the district in a quarter $t$ was in the bottom of top of the district ranking, instrumented by dummy variables equal to one if the district in a quarter $t$ was in the top (bottom) while very close to being out of it. For district is in the top, is defined as being in the $4^{th}$ or $5^{th}$ position. For districts in the bottom is defined as being in the $32^{th}$ or $33^{th}$ position. We control by District in the ranking threshold$_t$, defined as a dummy for quarters in which the district was close to the threshold for being in the top (bottom): between the $4^{th}$ and $7^{th}$ position for the top, and between the $29^{th}$ and $33^{th}$ position for the bottom. We include district and quarter fixed effects. Outcomes in the top of each column, measured in quarter $t+1$ in scale from 0 to 100. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean. Dep. Var* shows the average outcome in the districts outside the top and bottom positions in quarter $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Standard errors clustered by district in parentheses. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

**Table B10:** Intensity of exposure to flagging - heterogeneity by political alignment

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence$_{t+1}$ | | Student attendance$_{t+1}$ | | Functional facilities$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| # Times flagged$_t$ (z-score) | 0.067 | -0.021 | 0.438*** | -0.225 | 0.952*** | -0.151 |
| | (0.113) | (0.145) | (0.123) | (0.203) | (0.225) | (0.231) |
| Not fully aligned × # Times flagged$_t$ (z-score) | 0.071 | 0.087 | -0.241* | 0.074 | -0.035 | 0.053 |
| | (0.122) | (0.164) | (0.141) | (0.222) | (0.259) | (0.261) |
| Fully aligned × # Times flagged$_t$ (z-score) | 0.214* | 0.204 | -0.057 | 0.261 | 0.017 | 0.087 |
| | (0.119) | (0.162) | (0.136) | (0.217) | (0.258) | (0.262) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 |
| Mean Dep. Var. | 91.6 | 91.6 | 87.6 | 87.6 | 90.2 | 90.2 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 |
| 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 |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes: Notes:* The unit of analysis is the school-year. This table show the results from estimating equation 3 adding an additional interaction for the political alignment to test for heterogeneity of the main results of Table 4. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A report for the schools performing above the median of the markaz average performance. Panel B reports for schools performing below the median of the markaz average performance. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Following Callen, Gulzar and Rezaee (2020) we define: i) maraakiz fully aligned: all schools lie within constituencies where the winners are part of the chief minister party, ii) not fully aligned, and iii) not aligned: no constituency with the same party as the chief minister. Year is measured as school-year (September to May). *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Teacher presence and student attendance are measured as the percentage of present teachers/students. Functional facilities is measured as the percentage of the school's functional infrastructure. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

# C Tracing Impacts Through the Machinery of Government

**Table C1:** Intensity of exposure to flagging - effect on AEO's effort

| **Panel A: Bureaucratic oversight** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Flagging variable | Teacher presence | | Student attendance | | Functional facilities | | Math | | English | | Urdu | |
| | Dependent variable (range: 0-100) Visited schools$_{t+1}$ | | | | | | | | | | | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.721*** | 0.387 | 0.727*** | -0.206 | 0.012 | -0.506 | 0.205 | 1.518** | -0.386 | -0.594 | 0.125 | 0.002 |
| | (0.223) | (0.353) | (0.267) | (0.430) | (0.242) | (0.374) | (0.250) | (0.658) | (0.320) | (0.483) | (0.232) | (0.483) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 | 67,385 | 67,385 | 67,383 | 67,383 | 67,384 | 67,384 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. | 68.2 | 68.2 | 68.2 | 68.2 | 68.2 | 68.2 | 68.2 | 68.2 | 68.2 | 68.2 | 68.2 | 68.2 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 210.8 | | 258.6 | | 382.5 | | 34.8 | | 157.3 | | 53.3 |
| **Panel B: Transfers and postings** | | | | | | | | | | | | |
| Flagging variable | Teacher presence | | Student attendance | | Functional facilities | | Math | | English | | Urdu | |
| | Dependent variable (range: 0-100) Change head teacher$_{t+1}$ | | | | | | | | | | | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.246*** | 0.336** | -0.017 | -0.184 | 0.420*** | 0.235 | -0.109 | -0.028 | 0.019 | 0.105 | -0.020 | -0.184 |
| | (0.087) | (0.138) | (0.101) | (0.155) | (0.102) | (0.168) | (0.101) | (0.239) | (0.123) | (0.192) | (0.090) | (0.179) |
| N. of obs. | 257,592 | 257,592 | 257,865 | 257,865 | 254,058 | 254,058 | 67,385 | 67,385 | 67,383 | 67,383 | 67,384 | 67,384 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 |
| Mean # Times flagged | 0.87 | 0.87 | 1.59 | 1.59 | 1.61 | 1.61 | 0.038 | 0.038 | 0.34 | 0.34 | 0.065 | 0.065 |
| SD # Times flagged | 1.42 | 1.42 | 2.19 | 2.19 | 2.81 | 2.81 | 0.24 | 0.24 | 0.96 | 0.96 | 0.33 | 0.33 |
| First stage F-stat | | 210.8 | | 258.6 | | 382.5 | | 34.8 | | 157.3 | | 53.3 |
| Markaz FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| District time trends | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* The unit of analysis is the school-year. This table show the results from estimating equation 3 on bureaucratic effort outcomes as a measure of the machinery of government. Outcome in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A report on the bureucratic oversight dependent variable. Panel B reports on the transfers and posting dependent variable. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Change head teachers measure the percentage of months a school reported changed of head teacher. Visited schools measure the percentage of months where schools received a visit by a public official. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

**Table C2:** Intensity of exposure to flagging - effect on funds by category

**Panel A: School outcomes flagging**

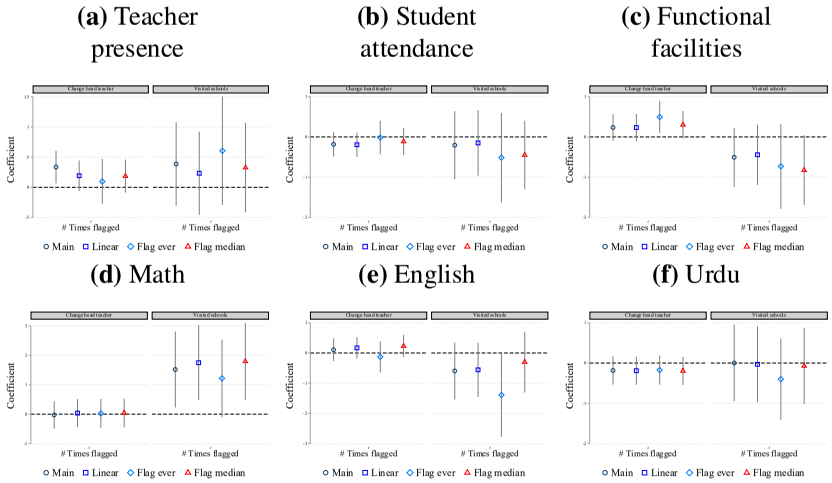
| Flagging variable | Teacher presence | | | | Student attendance | | | | Functional facilities | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dependent variables (logs) | | | | | | | | | | | |
| | Government Funds$_{t+1}$ | | Non Government Funds$_{t+1}$ | | Government Funds$_{t+1}$ | | Non Government Funds$_{t+1}$ | | Government Funds$_{t+1}$ | | Non Government Funds$_{t+1}$ | |
| | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.198*** | 0.140* | 0.063** | 0.007 | 0.202*** | 0.247** | -0.041 | -0.058 | -0.053 | -0.043 | -0.020 | 0.015 |
| | (0.047) | (0.081) | (0.027) | (0.046) | (0.053) | (0.102) | (0.032) | (0.055) | (0.045) | (0.094) | (0.033) | (0.052) |
| N. of obs. | 257,592 | 257,592 | 257,592 | 257,592 | 257,865 | 257,865 | 257,865 | 257,865 | 254,058 | 254,058 | 254,058 | 254,058 |
| Number markaz | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 | 3,567 |
| Mean Dep. Var. (unlogged) | 40,146 | 40,146 | 4,712 | 4,712 | 40,146 | 40,146 | 4,712 | 4,712 | 40,146 | 40,146 | 4,712 | 4,712 |
| Mean # Times flagged | 0.87 | 0.87 | 0.87 | 0.87 | 1.59 | 1.59 | 1.59 | 1.59 | 1.61 | 1.61 | 1.61 | 1.61 |
| SD # Times flagged | 1.42 | 1.42 | 1.42 | 1.42 | 2.19 | 2.19 | 2.19 | 2.19 | 2.81 | 2.81 | 2.81 | 2.81 |
| First stage F-stat | | 210.8 | | 210.8 | | 258.6 | | 258.6 | | 382.5 | | 382.5 |

**Panel B: Student scores flagging**

| Flagging variable | Math | | | | English | | | | Urdu | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dependent variables (range: 0-100) | | | | | | | | | | | |
| | Government Funds$_{t+1}$ | | Non Government Funds$_{t+1}$ | | Government Funds$_{t+1}$ | | Non Government Funds$_{t+1}$ | | Government Funds$_{t+1}$ | | Non Government Funds$_{t+1}$ | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| # Times flagged$_t$ (z-score) | 0.041* | 0.075 | 0.042** | 0.079 | -0.040 | -0.016 | -0.032 | -0.040 | 0.006 | 0.021 | 0.041* | 0.052 |
| | (0.024) | (0.056) | (0.021) | (0.062) | (0.025) | (0.043) | (0.027) | (0.040) | (0.021) | (0.047) | (0.023) | (0.044) |
| N. of obs. | 67,385 | 67,385 | 67,385 | 67,385 | 67,383 | 67,383 | 67,383 | 67,383 | 67,384 | 67,384 | 67,384 | 67,384 |
| Number markaz | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 | 2,721 |
| Mean Dep. Var. (unlogged) | 40,146 | 40,146 | 4,712 | 4,712 | 40,146 | 40,146 | 4,712 | 4,712 | 40,146 | 40,146 | 4,712 | 4,712 |
| Mean # Times flagged | 0.038 | 0.038 | 0.038 | 0.038 | 0.34 | 0.34 | 0.34 | 0.34 | 0.065 | 0.065 | 0.065 | 0.065 |
| SD # Times flagged | 0.24 | 0.24 | 0.24 | 0.24 | 0.96 | 0.96 | 0.96 | 0.96 | 0.33 | 0.33 | 0.33 | 0.33 |
| First stage F-stat | | 34.8 | | 34.8 | | 157.3 | | 157.3 | | 53.3 | | 53.3 |
| Markaz FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Time FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| District time trends | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| # Times in threshold$_t$ FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

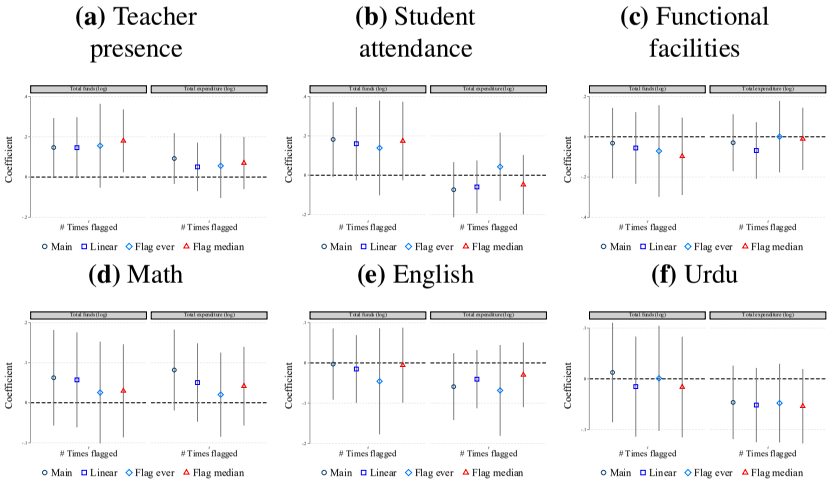
*Notes:* The unit of analysis is the school-year. This table show the results from estimating equation 3 on the dissagregation of total development funds to test the source of the effects observed in Table 5. Outcomes in the top of each column, measured in year $t+1$ in scale from 0 to 100. Panel A report on the schooling outcomes flagging. Panel B reports on the student scores flagging. # Times flagged$_t$ is the number of times a markaz was flagged in year $t$ in the outcome reported at the top of the column. # Times flagged threshold$_t$ is the number of times flagged while being close to the flagging threshold in the outcome reported at the top. # Times flagged is normalized (z-score). OLS columns show the results from equation 3. 2SLS columns show the results after instrumenting the # Times flagged by # Times flagged threshold. Year is measured as school-year (September to May). The first stage is estimated through equation 2. *First stage F-stat* show the Kleibergen and Paap (2006) F-statistic. *Mean # Times flagged* and *SD # Times flagged* indicate the mean and standard deviation of # Times flagged$_t$. *Mean. Dep. Var* shows the average outcome in the markaz in year $t$. Unlogged government and non government funds in pakistani rupees. Standard errors clustered by markaz in parentheses. $^{*}$ $p<0.10$, $^{**}$ $p<0.05$, $^{***}$ $p<0.01$.

**Figure C1:** Intensity of exposure to flagging - effect on effort - robustness by instrument definition




*Note:* This figure presents the results from estimating equation 3 on effort variables under different definitions of the instrument as a robustness test for the results in Table C1. The *Main* instrument is defined as dummy variables for each month of a year equal to one if in a such month a markaz was flagged while close to the threshold. The additional instrument definitions are measured as follows – *Linear*: normalized number of times flagged close to the threshold sample. *Flag Ever*: a dummy equal to one if the markaz reported to be flagged at least once in a year, and zero otherwise.*Flag Median*: a categorical variable dividing between markaz flagged above or below the median of the distribution. # Times flagged counts the number of times a markaz was flagged in year $t$ in the outcome reported in the title of the subgraph. The number of times flagged is normalized (z-score). Year is measured as school-year (September to May). Change head teachers measure the percentage of months a school reported changed of head teacher. Visited schools measure the percentage of months where schools received a visit by a public official. Error bars at the 95 percent level, clustered at the markaz level, are presented for each coefficient.

**Figure C2:** Intensity of exposure to flagging - effect on budget - robustness by instrument definition




*Note:* This figure presents the results from estimating equation 3 on budget variables under different definitions of the instrument as a robustness test for the results in Table 5. The *Main* instrument is defined as dummy variables for each month of a year equal to one if in a such month a markaz was flagged while close to the threshold. The additional instrument definitions are measured as follows – *Linear*: normalized number of times flagged close to the threshold sample. *Flag Ever*: a dummy equal to one if the markaz reported to be flagged at least once in a year, and zero otherwise.*Flag Median*: a categorical variable dividing between markaz flagged above or below the median of the distribution. # Times flagged counts the number of times a markaz was flagged in year $t$ in the outcome reported in the title of the subgraph. The number of times flagged is normalized (z-score). Year is measured as school-year (September to May). Total funds and total expenditure are log transformed. Error bars at the 95 percent level, clustered at the markaz level, are presented for each coefficient.

## D Naive Evaluations of Response

Building on the discussion in Section 6, we implemented a stacked difference-in-discontinuities analysis , where we compare schools in flagged and non-flagged maraakiz, before and after the flagging occurs. The difference-in-discontinuities allow us to compare maraakiz in the threshold sample, but we also report results for the full sample. The stacking allows us 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) and estimate features of the dynamics of schools in flagged and non-flagged maraakiz.

To explore the extent to which the lack of an effect is driven by a natural return of the outcomes to their original state, we estimate the following event study equation:

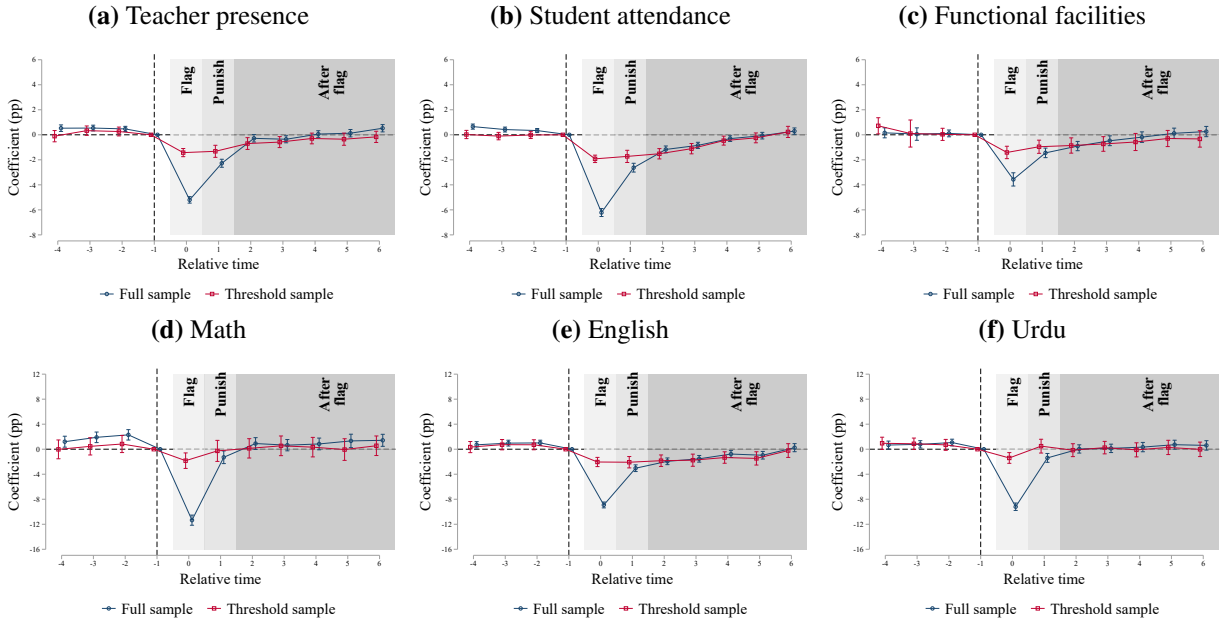
$$Y_{s,m,d,t,e} = \sum_{j \neq -1}^{J} \beta_j \cdot (T_{m,d,e} \times \mathbb{1}[j = t]) + \alpha_{m,d,e} + \lambda_{t,e} + \delta_{dt} + \varepsilon_{s,m,d,t,e} \tag{4}$$

where subscripts $s, m, d, t$ are for school, markaz, district, and time. All components are indexed at the flagging event panel $e$. $Y_{s,m,d,t,e}$ is the outcome for school $s$, within markaz $m$, in district $d$. $T_{m,d,e}$ equals 1 for schools in a flagged markaz $m$. $\beta_j$ captures the effect of being in a flagged markaz for each relative time $t = j$. $\alpha_{me}$ is for markaz fixed effects, and $\lambda_{te}$ is for time fixed effects. We also include $\delta_{dt}$ to absorb district linear time trends. $\varepsilon_{s,m,d,t,e}$ is the error term clustered at the markaz level. We stack for four pre-periods and seven post-periods.

Figure D1 reports the event studies for each outcome variable we study. The y-axis reports $\beta$ 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.

The recovery to pre-treatment means is some 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 D2 plots over time impacts on first-differenced outcomes that are reported above in Figure D1. 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

**Figure D1:** Event study - flagging effect on performance



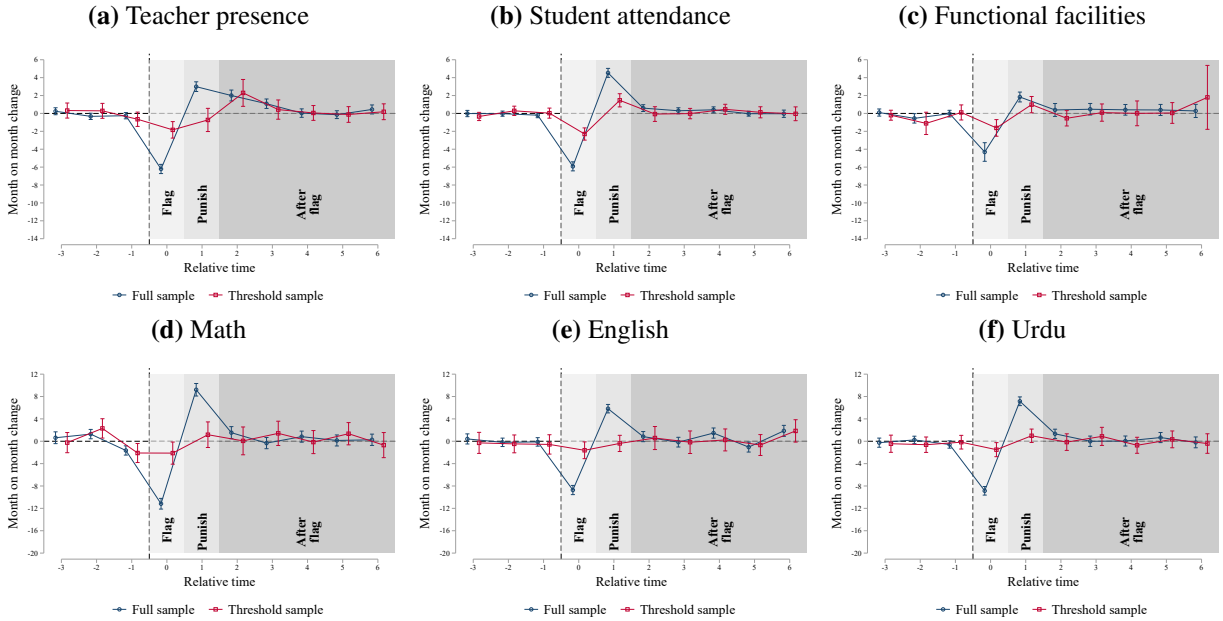

*Note:* This figure presents results from estimating event studies based on equation 4 using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through 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. 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. Scores variables are measured as the share of correct answers of students in standardized exams.

case that punishment 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 0.001). This shows that the rate at which teachers return to school 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, suggesting that punishment is not bringing any further improvement in the rate of recovery. The results for flagging on other outcomes are all indistinguishable from zero. Taken together, these results show that there is a small impact of command-and-control approaches that only occurs in the short-term on that margin of schooling most responsive to hierarchical pressure: personnel attendance. All other dynamics are equivalent to a reversion to the mean.

Table D1 show the average coefficients of the event study of Figure D1. The first column for

**Figure D2:** Punishment Period vs Reversion to Mean - Month on Month Changes




*Note:* This figure presents results from estimating month-by-month coefficients based on equation 4 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. 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. Scores variables are measured as the share of correct answers of students in standardized exams. 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) –0.001–. Panel (b) –0.99–. Panel (c) –0.99–. Panel (d) –0.75–. Panel (e) –0.19–. Panel (f) –0.83–.

each variable reports the full sample, and the second shows the threshold sample. Panel A reports outcomes relating to school functioning. Coefficients for *Flag* and *Punish* represent the first negative shock and the immediate recovery. The coefficients for the *After flag* report the trend after the immediate recovery. The coefficients are small compared to the mean of the dependent variable, but negative. Panel B of Table D1 presents the results for the student test score variables. We observe the same pattern of results as in Panel A. The results imply that the oversight scheme had negligible impacts on school functioning nor student outcomes, but rather that flagged and non-flagged schools facing a similar shock returned to equilibria. As an alternative specification for the threshold sample, we implement a regression discontinuity design for the effects in the *Punish* period around the flagging threshold. For all outcomes, the effects align with those of Table D1 (results available on request).

**Table D1:** Monthly monitoring effect on performance - markaz flagging

**Panel A: School outcomes**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence | | Student attendance | | Functional facilities | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| T×Flag | -5.593*** | -1.566*** | -6.603*** | -1.971*** | -3.658*** | -1.630*** |
| | (0.132) | (0.153) | (0.164) | (0.143) | (0.275) | (0.290) |
| T×Punish | -2.668*** | -1.471*** | -3.031*** | -1.804*** | -1.542*** | -1.175*** |
| | (0.153) | (0.213) | (0.178) | (0.251) | (0.197) | (0.301) |
| T×After flag | -0.403*** | -0.606*** | -0.889*** | -0.810*** | -0.356** | -0.787** |
| | (0.092) | (0.139) | (0.080) | (0.131) | (0.163) | (0.305) |
| Sample | Full | Threshold | Full | Threshold | Full | Threshold |
| N. of obs. | 8,202,224 | 673,614 | 6,080,752 | 693,441 | 8,737,264 | 1,085,311 |
| Mean Dep. Var. before | 92.9 | 87.9 | 91.7 | 87.2 | 97.2 | 95.6 |

**Panel B: Students scores**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Math | | English | | Urdu | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| T×Flag | -12.710*** | -2.158*** | -9.599*** | -2.511*** | -9.826*** | -2.014*** |
| | (0.338) | (0.512) | (0.196) | (0.265) | (0.250) | (0.348) |
| T×Punish | -2.654*** | -0.595 | -3.732*** | -2.556*** | -2.018*** | -0.121 |
| | (0.416) | (0.741) | (0.238) | (0.393) | (0.323) | (0.503) |
| T×After flag | -0.342 | -0.034 | -1.749*** | -1.851*** | -0.273 | -0.551* |
| | (0.258) | (0.469) | (0.186) | (0.268) | (0.203) | (0.326) |
| Sample | Full | Threshold | Full | Threshold | Full | Threshold |
| N. of obs. | 2,749,969 | 78,564 | 1,017,291 | 202,812 | 2,461,051 | 142,267 |
| Mean Dep. Var. before | 87.0 | 73.6 | 78.0 | 70.9 | 84.6 | 72.7 |
| 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 4. 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. 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. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

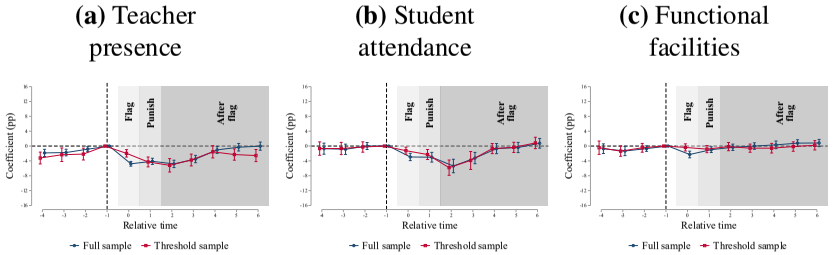
**Robustness to difference-in-discontinuities approach** A concern of our specification is that the bureaucrats' responses might have happened in expectation of the scheme implementation. We test whether the introduction of accountability under command-and-control management created significant educational outcomes changes by estimating the first flag's effect. As such, first-time flagged AEOs should have the highest immediate incentives to avoid punishment. The event study of Figure D3 shows no significant improvement, suggesting no relevant bureaucratic responses appeared from the immediate implementation of accountability in the command-and-control scheme.

We test for fewer post-periods in the stacked dataset to explore the sensibility of the results to the data structure. Figure D4 shows that the patterns remain consistent with a reversion to the mean. We estimate the effects for orange flagging in Figure D5 to validate that no effects are perceived in higher flagging thresholds. Figure D6 test for flagging at the tehsil level to explore effects on a more aggregate administrative unit. We find no significant effects from more aggregated flagging measures.

We further explore the comparison of maraakiz with the same flagging path before the negative shock to account for selection from maraakiz constantly flagged given the high frequency of the scheme. We do so by re-building the stacked dataset but allowing flagging in the pre-periods. We then build a 'flagging history FE' indicator such that we compare maraakiz that, before the relevant event-flagging, experienced exactly the same flagging behaviour. Table D2 shows the results. The coefficients of *After flag* are positive, but, when comparing it against the coefficients for flag *T*, the total effects result in small and negative, consistent with our previous results.

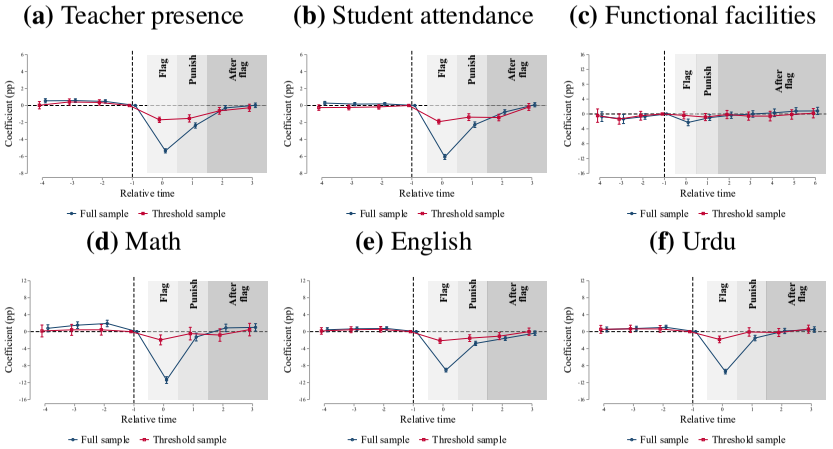
Finally, we also test for alternative difference-in-differences estimators. Figure D7 shows the results for the Sun and Abraham (2021) estimator, assuming markaz remain 'treated' after the first time flagged. Figure D8 implements the $\text{DiD}_l$ estimator by De Chaisemartin and D'Haultfoeuille (2022) to account for the turning on/off of the treatment. The results follow the same patterns as those of the stacked design, suggesting that flagged units follow a reversion to the mean.

**Figure D3:** Event study - first time flagging effect on performance - flagged units




*Note:* This figure presents results from estimating event studies based on equation 4 using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The effects are estimated 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. 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. *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 D4:** Event study - flagging effect on performance short stack




*Note:* This figure presents results from estimating event studies based on equation 4 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. 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. Scores are measured as the percentage of correct answers in standardized tests. *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.

**Table D2:** Monitoring effect on performance - markaz flagging - flagging history FE

**Panel A: School outcomes**

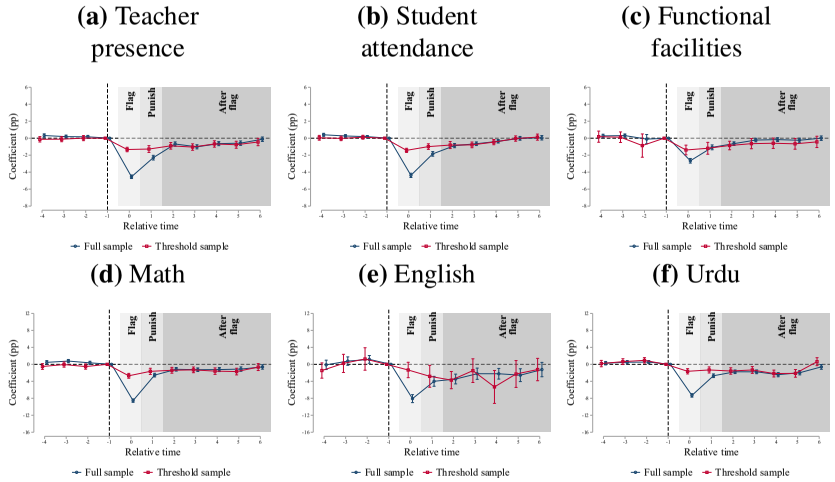
| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Teacher presence | | Student attendance | | Functional facilities | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| T | -2.544*** | -1.123*** | -4.624*** | -2.587*** | -9.128*** | -2.777*** |
| | (0.103) | (0.101) | (0.149) | (0.133) | (0.276) | (0.117) |
| T×Flag | -3.919*** | -1.000*** | -3.130*** | -0.034 | 1.087*** | 0.618*** |
| | (0.106) | (0.120) | (0.135) | (0.133) | (0.121) | (0.109) |
| T×Punish | -0.979*** | -0.707*** | -0.290** | -0.051 | 2.205*** | 0.333*** |
| | (0.100) | (0.135) | (0.130) | (0.156) | (0.141) | (0.115) |
| T×After flag | 1.512*** | 0.386*** | 2.927*** | 1.228*** | 4.890*** | 1.280*** |
| | (0.111) | (0.111) | (0.171) | (0.151) | (0.225) | (0.119) |
| Sample | Full | Threshold | Full | Threshold | Full | Threshold |
| N. of obs. | 12,160,124 | 2,332,142 | 11,281,261 | 2,946,436 | 13,858,866 | 2,440,073 |
| Mean Dep. Var. before | 91.5 | 87.4 | 88.3 | 86.3 | 93.0 | 91.1 |

**Panel B: Students scores**

| | Dependent variables (range: 0-100) | | | | | |
|---|---|---|---|---|---|---|
| | Math | | English | | Urdu | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| T | -3.443*** | -0.987** | -4.152*** | -1.988*** | -3.553*** | -1.117*** |
| | (0.329) | (0.406) | (0.156) | (0.171) | (0.240) | (0.269) |
| T×Flag | -11.897*** | -1.928*** | -7.123*** | -2.110*** | -9.262*** | -1.745*** |
| | (0.359) | (0.533) | (0.154) | (0.201) | (0.238) | (0.333) |
| T×Punish | -1.952*** | -0.428 | -1.758*** | -1.683*** | -1.493*** | -0.277 |
| | (0.404) | (0.644) | (0.151) | (0.235) | (0.289) | (0.458) |
| T×After flag | 1.634*** | 0.485 | 1.601*** | 0.077 | 1.765*** | 0.396 |
| | (0.323) | (0.522) | (0.168) | (0.203) | (0.226) | (0.340) |
| Sample | Full | Threshold | Full | Threshold | Full | Threshold |
| N. of obs. | 2,871,621 | 97,336 | 1,999,008 | 737,392 | 2,666,695 | 246,385 |
| Mean Dep. Var. before | 86.7 | 73.0 | 74.7 | 70.4 | 84.2 | 74.8 |
| 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 |

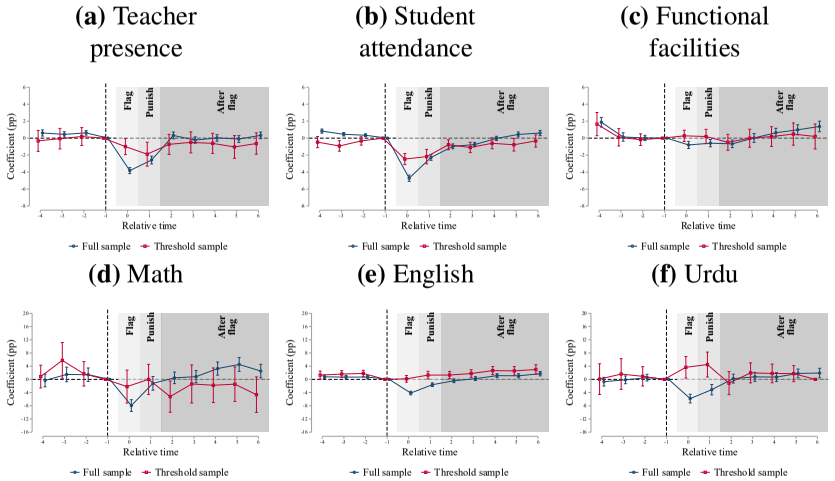
*Notes:* Results from estimating a modified version of equation 4, including flagging history FE instead of markaz FE. Flagging history is built from concatenating the flagging status in the three periods before the observed flagging. The specification compares maraakiz that had the same flagging path before the negative shock. Flagging history is not a markaz attribute, so the term $T_{mde}$ from equation 4 is not absorbed and the interactions can be compared against it. 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. 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 D5:** Event study flagging effect on performance orange threshold



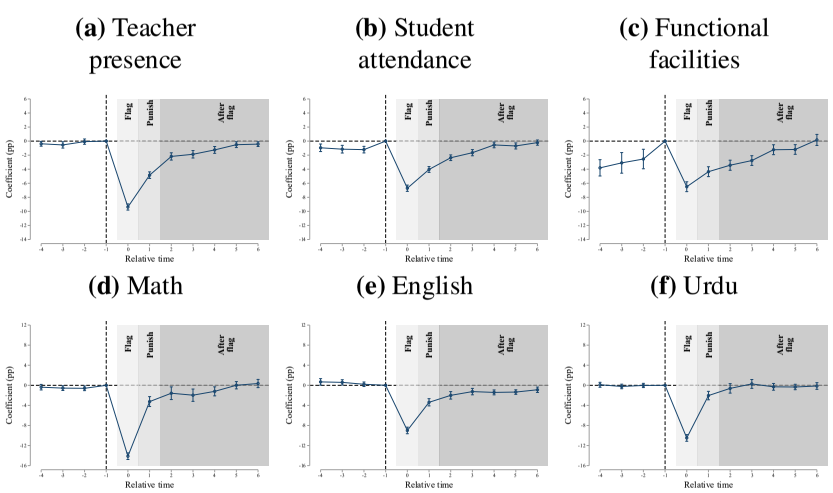

*Note:* This figure presents results from estimating event studies based on equation 4 using -1 as the base period, comparing schools in orange-flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained 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.

**Figure D6:** Event study - flagging effect on performance tehsil-wing flagging



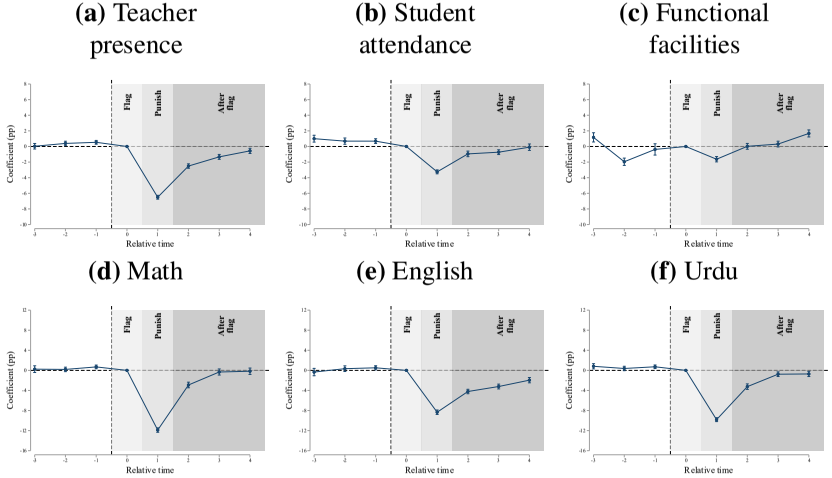

*Note:* This figure presents results from estimating event studies based on equation 4 using -1 as the base period, comparing schools in flagged and non-flagged tehsil-wing. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through 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.

**Figure D7:** 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. 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. Scores are measured as the percentage of correct answers in standardized tests. *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 D8:** 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. 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. Scores are measured as the percentage of correct answers in standardized tests. *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.