The Limits of Centralized Accountability in the Public Sector*

Saad Gulzar Juan Felipe Ladino Muhammad Zia Mehmood Daniel Rogger†

November 11, 2026

A long-established approach to government management combines the transmission of information up administrative hierarchies with centralized decision-making and accountability, an approach commonly described as 'command and control'. While governments in over 80 countries have adopted such initiatives, evidence on the efficacy of the accountability pressures they generate remains thin. We study a large-scale implementation of this approach in public schools across Punjab, Pakistan, where poorly performing jurisdictions were automatically flagged for the attention of senior management. Exploiting quasi-random variation in the intensity of accountability across 52,000 schools, we show that the resulting de facto punishments had negligible effects on school and student outcomes. Nor did they generate observable administrative responses across the service-delivery chain, including school visits, personnel transfers, and budget allocations. Yet qualitative evidence indicates bureaucrats believed that the system was effective. We show that this perception is consistent with a simple before–after comparison indicating improvements being attributed to intervention by the accountability system, when they instead reflected mean reversion following transient, idiosyncratic shocks. Treating these naïve comparisons as policymakers' prior beliefs in a Bayesian framework yields a posterior distribution close to our causal estimates near zero. These findings show how centralized accountability can appear effective even when it generates little sustained organizational response, raising questions about its effectiveness as a foundation for command and control management in complex service-delivery environments.

*We thank Faisal Bari, Michael Callen, Jishnu Das, Alessandra Fenizia, Koen Geven, Dan Honig, Clare Leaver, Rabea Malik, Imran Rasul, and Martin Williams for their useful comments; and seminar participants at Berkeley, the Education Commission, Georgetown, the Institute of Development and Economic Alternatives, the Institute for Fiscal Studies, Oxford, and the World Bank. We thank Belen Torino for research assistance and the Punjab Program Monitoring and Implementation Unit for providing us with data and institutional details. 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. We 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.

†Gulzar: Department of Political Science and Keough School of Global Affairs, University of Notre Dame; Ladino: Department of Economics, Stockholm University; Mehmood: Department of Economics, Occidental College; Rogger: World Bank Development Impact Evaluation Research Department.

1 Introduction

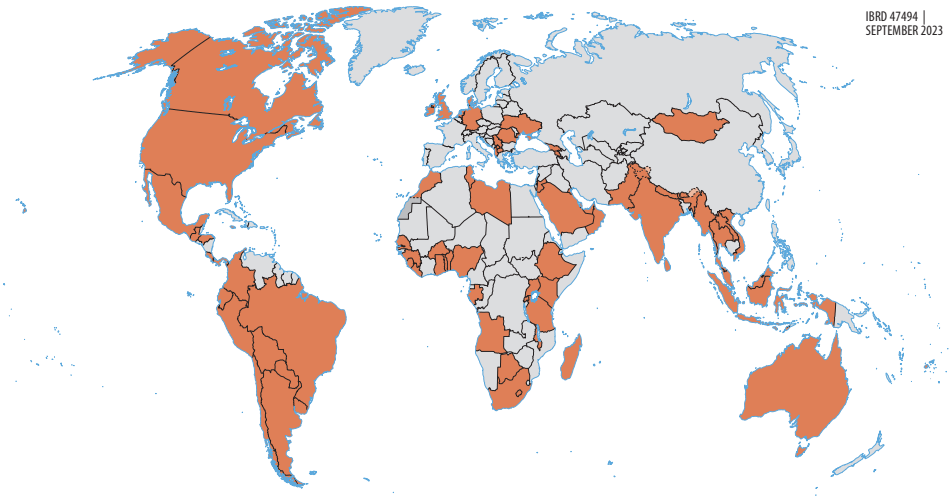
How the bureaucracy performs is fundamental to the provision of high-quality public services to ordinary citizens around the world (Wilson, 1989; Simon, 1997; Finan, Olken and Pande, 2017; Pollitt and Bouckaert, 2017; Besley et al., 2022; Brierley et al., 2023). The vast majority of reforms to government administration implemented at scale relate to shaping the de facto incentives in the bureaucracy instead of introducing changes in legal and fiscal environments. A canonical de facto bureaucratic reform is command and control management: a system in which targets are set centrally and imposed throughout the hierarchy, performance is strictly monitored, and accountability is enforced through regular top-down sanctioning routines. 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 its efficacy. Following the purported success of British Prime Minister Tony Blair's 'delivery unit', over 150 documented instances of command and control delivery approaches have been implemented across more than 80 countries, covering roughly 40% of countries worldwide (see Figure 1). Adoption has accelerated in recent years, with more than half of these systems established after 2010, and nearly half including education as a focus sector. These programs are resource-intensive and widespread with over half receiving external funding and nearly 70% relying on external technical assistance, often involving the creation of new bureaucratic structures (Mansoor et al., 2021). Despite their rapid diffusion and substantial resource requirements, evidence on these reforms, and in particular on the incentive effects of their accountability components, remains thin.

We study a system-wide command and control delivery approach scheme in the public education administration of Punjab, Pakistan, one of the largest education bureaucracies in the world. Monthly education data from over 50 thousand primary and middle 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 the most senior members of the executive (Barber, 2013; Chaudhry and Tajwar, 2020; Malik and Bari, 2023).

Our analysis focuses on how the intensity of accountability implemented within the scheme affects administrative actions and educational outcomes. We collect administrative data on all 52,000 primary and middle public schools in Punjab from December 2011 to May 2018 on which the scheme

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



Source: Map produced by the World Bank using data from Mansoor et al. (2021).

was built and digitize the monthly reports created for senior managers that flagged underperforming administrative units. The monitoring reports present performance metrics drawn from these data, aggregated at the administrative unit, for a range of school outcomes including teacher presence, student attendance, functional facilities, and from January 2016, student scores on standardized tests. 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 progression 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 us to estimate even small effects with precision, providing an unusually rich picture of the reform.

We use a regression discontinuity aggregation (RDA) design to estimate how the intensity of accountability in this command and control system affects subsequent school performance. Each month, a jurisdiction (*markaz*, plural *maraakiz*) is flagged if a sufficient share of its schools falls below a fixed performance threshold on the outcome of interest. Because chronically underperforming *maraakiz* are mechanically flagged more often, the RDA approach instruments a *markaz*'s annual flag count with the share of months in which it was flagged narrowly below the threshold. Under a local randomization assumption, these threshold-crossing months supply as-if random variation in flagging intensity, aggregated from monthly regression discontinuity events to the *markaz*-year level.

We find precisely estimated evidence that the intensity of accountability in the command and con-

trol scheme had no substantive impact on targeted school or student outcomes: teacher and student attendance, functional school facilities, and English, Mathematics, and Urdu test scores. For instance, the RDA analysis shows that a one standard deviation increase in annual flagging, which more than doubles mean flagging, improves subsequent teacher presence, an outcome clearly within the authority of public managers, by less than a tenth of a percentage point. Point estimates for the other outcomes are similarly below, or close to a percentage point, with standard errors below a third of a percentage point. We also show that the null extends to the underlying machinery of public education bureaucracy: flagged *maraakiz* see no additional bureaucratic visits, no rotation of head teachers, no change in non-salary budget allocations, and only minor increases in expenditures on school operations.

Despite these null effects, the program ran for six years. We present qualitative and quantitative evidence that suggests that this could be due to naive before-after comparison making the flag-triggered accountability response appear effective. Contemporaneous accounts from key actors whose support sustained the reform establish that they believed it was producing real improvements. However, zooming in to monthly data, we show that flagged jurisdictions return to pre-shock levels of service delivery on the same trajectory as non-flagged jurisdictions, following a classic reversion-to-the-mean pattern (Chay, McEwan and Urquiola, 2005). Without an appropriate counterfactual, managers could not separate this reversion from the effect of their own interventions. We test this interpretation in a Bayesian analysis that treats the naive estimates as senior managers' priors. Updating these priors with our causal estimates yields posteriors that concentrate near zero across all outcomes. Without this update, the belief that the scheme was working persisted.

A large literature emphasizes that knowledge-intensive public services rely on professional autonomy rather than tight hierarchical control. Work on professional organizations highlights the centrality of expert judgment and discretion (Mintzberg, 1979; Deci and Ryan, 1985; Freidson, 2001; Prendergast, 2002; Bechky and Chung, 2018; Honig, 2024; Ali et al., 2021; Deserranno et al., 2025; Cilliers and Habyarimana, 2023), while studies of street-level bureaucracy show that frontline actors must adapt to complex, context-specific environments (Lipsky, 1980; Mangla, 2022). A key mediator in the activation of autonomy is the degree of control a manager attempts to exert over their employees (Hood, 1991; Bloom and Van Reenen, 2010; Bloom et al., 2015; Rasul and Rogger, 2018; Rasul, Rogger and Williams, 2021; Banerjee et al., 2021; Ali et al., 2021; Carreri, 2021; Honig, 2021). Evidence on the impact of control mechanisms on public sector performance is mixed, with generally positive results from experiments that specify a narrow frontline task and tightly control the incentive provided (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 that attempt to influence administrators'

broader motivation and performance, or in those cases dealing with organizational dynamics (Falk and Kosfeld, 2006; Dickinson and Villeval, 2008; Bandiera et al., 2021; Muralidharan and Singh, 2020; Dodge et al., 2025; Mangla, 2022).

Wilson's classic categorization of public agencies (Wilson, 1989) identified schools as 'coping organizations' in which central managers "can observe neither the outputs nor the outcomes of their key operators" taken to cope with complex problems. Wilson argued that top-down management fails in these environments, and scholarship has since outlined mechanisms for these failures from inappropriate conceptions of the problems to delivering better services (Power, 1999; Scott, 1998; Eyben, 2010), gaming of associated targets (Kerr, 1995; Behn, 2003; Hood, 2006; Espeland and Sauder, 2007; Muller, 2018) and an undermining of local problem-solving systems (Argyris and Schön, 1997; McCubbins, Noll and Weingast, 1987; Mansbridge, 2014; Deserranno, Kastrau and León-Ciliotta, 2025). In the context of continued and resurgent interest in centralized approaches to public management (Hood, 1991; Kettl, 2005; Muralidharan and Sundararaman, 2011; Dal Bó, Finan and Rossi, 2013; Ashraf, Bandiera and Jack, 2014; Khan, Khwaja and Olken, 2019; Bertrand et al., 2020; Leaver et al., 2021), this paper uses the same data leveraged for accountability in centralized control to illustrate its failure to strengthen government performance in a key sector. As such, it provides some of the first at-scale evaluation of de facto, sanctions-based, control schemes and shows that corresponding pressure from managers does not engender substantial responses from public officials, however salient senior management makes this form of incentive provision. In Wilson's terminology, our findings cast doubt on efforts to transform education systems from coping organizations into 'production organizations' in which there is greater observability of outputs and outcomes. They suggest that the rapid expansion of digital technologies does not diminish the importance of the structural features of public management that govern how bureaucracies function.

The paper proceeds as follows: Section 2 describes the setting of the public service we study and describes the centralized monitoring scheme. Section 3 presents our core results on the impact of flagging. Section 4 explores the extent to which the scheme's persistence can be explained by a naive evaluation of the bureaucratic response. Section 5 showcases how managers should have updated their beliefs in response to appropriate counterfactual analysis. Finally, Section 6 provides discussion.

2 Public Education in Punjab

Home to roughly 128 million people, Punjab is the most populous province of Pakistan. Approximately 36 million of these residents are school-aged (between 5 and 16 years of age), with over 10.5 million enrolled in public schools (Pakistan Bureau of Statistics, 2023). The School Education Department is responsible for organizing and overseeing the education sector's performance in the province. With around 50,000 public schools employing roughly 320,000 teachers, the scale of managing public education in the province is substantial (PMIU-PESRP, 2025).

2.1 Education Implementation and Monitoring

The province's education system reflects layers of hierarchy common in large public bureaucracies. Each of the 36 districts has a district education authority, led by an Executive District Officer and three District Education Officers, that reports directly to the School Education Department. Districts are subdivided into administrative units called *tehsils*, which are further subdivided into areas of responsibility called *maraakiz* (plural of *markaz*). Below the district leadership team, the hierarchy consists of officers for each *tehsil* and Assistant Education Officers (AEOs) for each *markaz*, sitting directly above school principals. This structure completes a multi-link chain of command from the Executive to the school level. On average, a district-level manager tracks 192 *maraakiz*, and each AEO manages 20 schools.

Each layer is expected to manage the officers below in terms of motivating, monitoring, and punishing subordinates. AEOs are the frontline supervisory staff of this chain. Their formal duties include conducting inspection visits, observing and strengthening teaching practices, monitoring teacher and student attendance, assessing the functionality of school facilities, coordinating with head teachers, and ensuring the implementation of departmental policies (Malik and Bari, 2023). They communicate regularly with their supervising Deputy District Education Officer, submitting visit plans and reporting on school conditions, and they serve as the primary conduit through which senior management's instructions reach individual schools (Malik and Bari, 2023).

AEOs have the authority and capacity to affect change in schools. AEOs can visit schools, motivate and reprimand teachers, direct head teachers to correct specific problems, and follow up on whether flagged deficiencies have been addressed. They can also recommend formal disciplinary action to higher levels of the district and provincial hierarchy. However, final decisions over such action, as well as transfers and postings and development-budget allocation or execution, generally require approval or action by the Executive District Officer or provincial authorities rather than unilateral

action by an AEO.

Geographic dispersion and administrative complexity limit senior officials' ability to directly observe frontline performance, and this hierarchy is structured to make such management practical. Monitoring systems aim to relax these constraints by lowering the cost of aggregating information across administrative units, enabling centralized oversight of service delivery at scale.

2.2 Centralized Accountability Intervention

Datapacks An independent institution, the Program Monitoring and Implementation Unit (PMIU), is tasked with monitoring the performance of the district education administration. To this end, Monitoring and Evaluation Assistants, who report to PMIU, conduct randomized monthly inspection visits to public schools to collect information on key aspects of the school environment.

Beginning in December 2011, PMIU monitoring data were aggregated into standardized monthly performance reports known as 'datapacks.' Datapacks, which were the key feature of the command and control accountability we study here, summarized important indicators of school performance, including teacher presence, student attendance, and the functionality of school infrastructure (electricity, drinking water, toilets, and boundary walls). From 2016 onward, datapacks also reported results from standardized monthly assessments in Math, English, and Urdu administered to six randomly selected Grade 3 students during school inspection visits.

Performance on each indicator was benchmarked against fixed thresholds determined by the Chief Minister's team and presented using a color-coded system. Jurisdictions were classified as green, orange, or red, with red being the primary flag for underperformance relative to the target threshold. Our study period spans from the introduction of datapacks in December 2011 to May 2018, immediately prior to the national elections that led to a change in administration.

Datapacks for Centralized Accountability Building on the availability of high-frequency performance indicators generated through PMIU monitoring, the provincial government introduced a centralized accountability reform designed to transmit performance information through the administrative hierarchy and strengthen oversight of the education bureaucracy. Senior political leadership, led by the Chief Minister, oversaw the reform through a dedicated oversight committee supported by McKinsey & Company and external technical advisors. The system was structured as a top-down management framework in which key performance thresholds and priorities were determined centrally and communicated throughout the administrative hierarchy. As described by

Barber (2013), the lead technical advisor to the government on the project, the reform emphasized rapid adoption of this framework rather than iterative consensus-building across administrative layers:

> "Management theory might suggest that the Executive District Officers should have been consulted about these targets and trajectories. We judged otherwise ... Too often the pursuit of 'buy-in' from 'key stakeholders' becomes an excuse for slower, shallower action and sometimes for inaction." (Barber, 2013, p. 41)

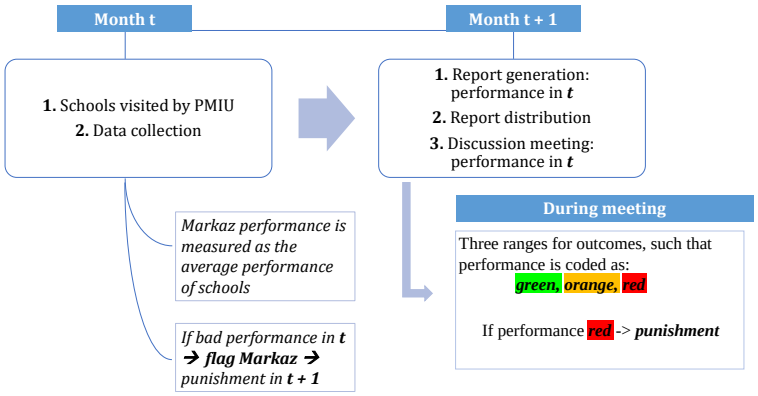
Regular review meetings were organized around the datapack reporting cycle, with monthly meetings at the district level and quarterly meetings at the provincial level providing structured opportunities for senior managers to assess performance of subordinate administrative units and initiate corrective actions. The color-coded flagging was a central focus of these meetings. A senior policy advisor noted: "we used to produce heatmaps in which we used to show [the Punjab city of] Bahawalpur as red and [the Punjab city of] Rahim Yar Khan as green, and suddenly the whole onus was on why Bahawalpur was still red and why have they not followed up on teacher and student absenteeism and basic facilities" (Malik and Bari, 2023, p. 24). A district official further reflected: "[Chief Minister] Shahbaz Sharif used to ask very difficult questions... Only the red [underperforming districts] were reprimanded, and [the green top-performing districts] were appreciated. The ones in amber were left alone ... He would use really, really strong words for those in the red zone." (Malik and Bari, 2023, p. 32). These accounts indicate that discussions and reprimands were triggered by flag status rather than marginal differences in the underlying indicators.

Figure 2(a) illustrates the design of the monitoring and accountability system. Data on all schools in the province were collected in month t through PMIU inspections and aggregated into *markaz*-level performance indicators, which were presented to district managers in month $t+1$. *Maraakiz* falling below fixed performance thresholds–specifically those flagged in red–became the focus of subsequent review discussions and reprimands.

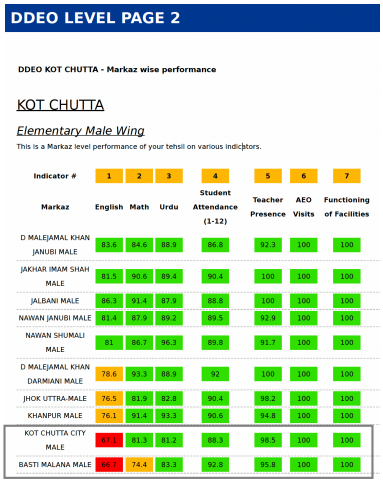
Figure 2(b) illustrates an example of a report used in the monthly meetings. It shows *maraakiz* compared in various performance indicators (in columns). In this example, the English test scores in 'Kot Chutta City Male' *markaz* and 'Basti Malana Male' *markaz* were highlighted red for falling below the flagging threshold. Figure 2(c) 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.

Figure 2: Details of the command and control structure

(a) Monitoring scheme

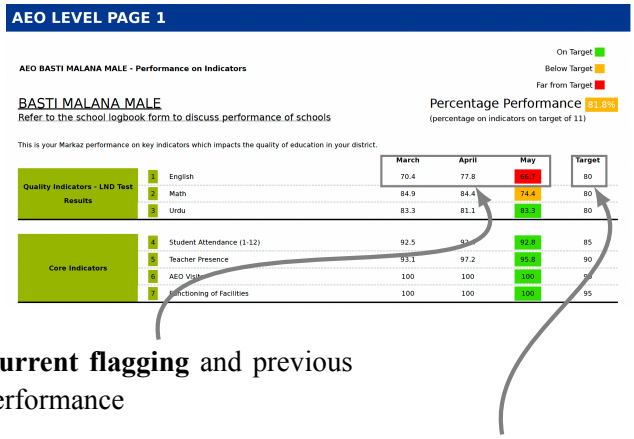


(b) Performance report by maraakiz



Flagged red **AEO censored**

(c) Details on current and previous maraakiz performance



Current flagging and previous performance

AEO not meeting target consistently

Notes: **(a)**, In month t schools are visited by Program Monitoring and Implementation Unit (PMIU) officers. Starting month $t+1$ this data is used to build *markaz* level performance reports, which are presented in the meetings. If the *markaz* average performance lies below the red flagging threshold, Assistant Education Officers (AEOs) are censured and punished. **(b)**, Example of a monthly performance report for the 'Kot Chutta' District. The rows are *maraakiz* within the district. The columns are performance indicators highlighted with the color of the flagging level. As an example, *maraakiz* 'Kot Chutta City Male' and 'Basti Malana Male' are flagged red in English scores performance. **(c)**, AEO-specific performance report for the *markaz* 'Basti Malana Male'. The report shows the time series for the evolution of performance indicators. Columns are for months. Rows are for indicators. The last month indicators are highlighted with the color of the flagging level. The last column indicates the expected target. For the English indicator, the *markaz* is always below the expected target.

Driven by strong and sustained commitment from senior political leadership, the centralized accountability system was implemented across the province and remained fully in operation throughout our study period. Supplementary Information A.3 provides additional evidence on the high level of implementation fidelity, documenting the regular production of datapacks and the contin-

ued use of performance indicators in review meetings across administrative tiers.

The combination of standardized performance indicators, hierarchical review structures, fixed performance thresholds, and sustained implementation over multiple years create a large-scale setting in which bureaucratic performance was managed through command and control accountability system originating from the top and operating across multiple levels of the administrative hierarchy. The breadth of administrative coverage, consistency and quality of data collection, and duration of the reform make this a particularly well suited setting to study how centralized, data-driven accountability systems function in a complex public-sector bureaucracy.

3 Intensity of Centralized Accountability

We study the impact of flagging accountability in the command and control approach on public education performance using variation in the number of times a *markaz* was flagged during the last year. Administrators of repeatedly flagged *maraakiz* face stronger de-facto punishments, and should therefore take actions that accumulate into better school outcomes. A school's performance in a given year thus depends on the extent to which its administrators effectively respond to this more intense flagging.

We aggregate the monthly datapacks to the school-year level and count, for each school, how many times its *markaz* was flagged. This count is not randomly assigned: systematically underperforming *maraakiz* are flagged more often, confounding any naive comparison. We address this identification concern with a regression discontinuity aggregation (RDA) design, which is formalized by Borusyak and Kolerman-Shemer (2024) and commonly used in settings where the treatment and outcome of interest are measured at a more aggregated level than the individual discontinuities that generate the causal variation. In our case, the treatment is annual flagging intensity, but each monthly flag-or-not decision is its own regression discontinuity event, governed by whether the *markaz*'s running variable crosses the threshold that month. RDA recovers the causal effect of annual flagging intensity by pooling the quasi-random variation across these monthly discontinuities. Concretely, we instrument a *markaz*'s total annual flag count with the share of its months flagged just below the threshold, the months in which the flagging decision was effectively a coin flip. See the Methods Section 7.2 for further details.

Our results suggest that more intense exposure to accountability does not substantially improve average school performance. Figure 3(a) plots the estimated effects of flagging intensity on the next year's average school outcomes and student scores, separately for ordinary least squares (OLS) and RDA, where the RDA estimates capture the causal effect of more intense flagging. Coefficients

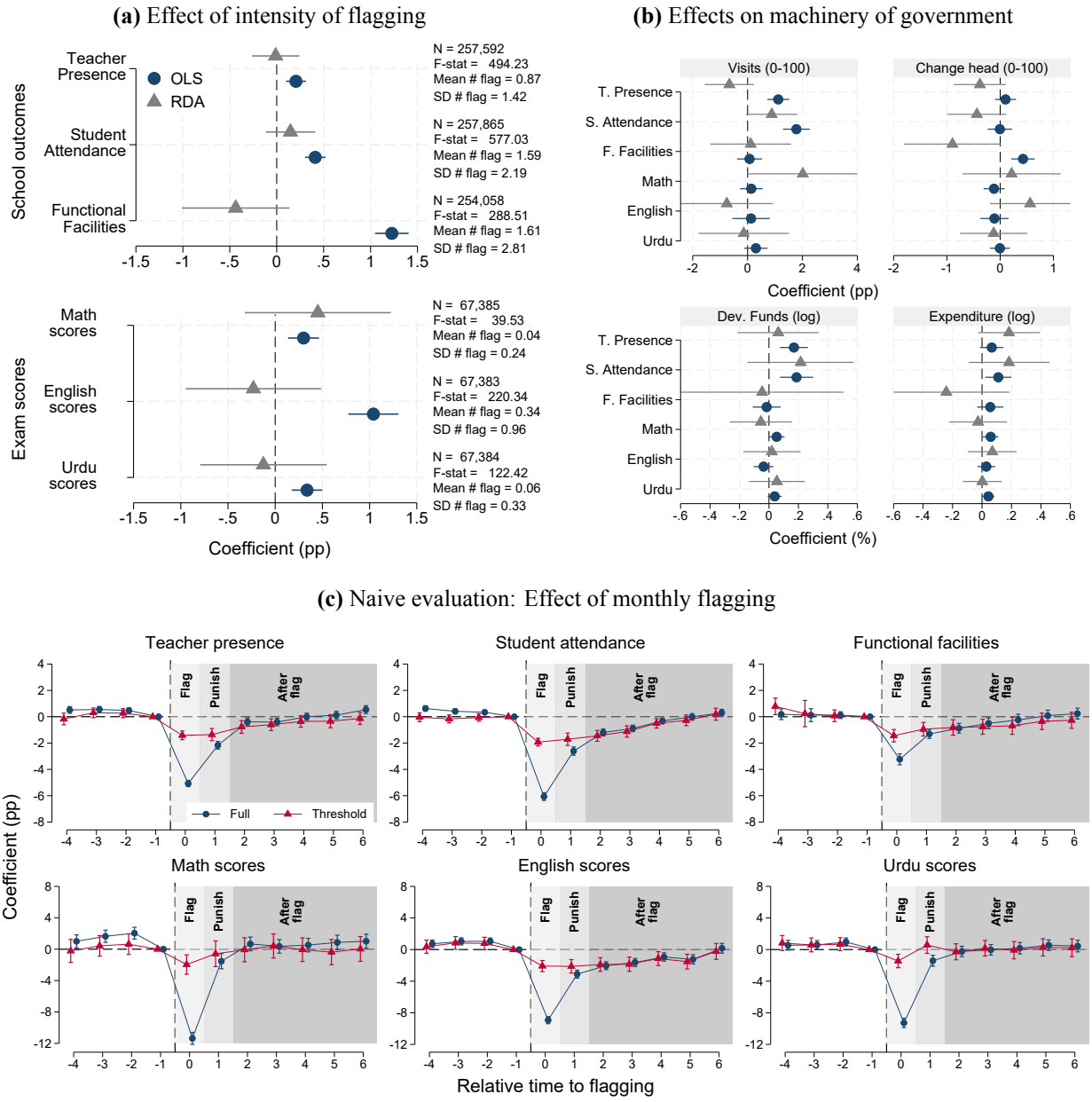
are expressed in percentage points per one-standard-deviation increase in flagging intensity. A standard-deviation increase represents roughly 0.24 to 2.81 additional flags per year across indicators.

We find that across outcomes from teacher and student attendance, functional facilities, to student test scores, the impacts of more intense command and control accountability are small and tightly estimated, ruling out anything beyond economically negligible effects. All RDA coefficients are below one percentage point in absolute value and statistically indistinguishable from zero ($p > 0.1$). The 95% confidence intervals exclude improvements larger than half a percentage point for school outcomes and 1.2 percentage points for test scores (Supplementary Table C1). OLS estimates are all below or close to one percentage point. We present alternative specifications in Supplementary Information C.1.

These null effects are robust and extend across alternative margins along which more intense accountability within the scheme might plausibly have produced a response (see Supplementary Information C.3 for details): when we examine the best- and worst-performance within a year; when we test for heterogeneous responses across school characteristics that might reflect administrator targeting; at the green/orange reward threshold instead of the orange/red threshold; when we allow for differential effects at the district level, both by interacting flagging intensity with indicators for bottom- and top-five district rankings and by treating district rank itself as the treatment in a district-by-quarter specification; and when we test for differential effects by political alignment with the Chief Minister's party (Callen, Gulzar and Rezaee, 2020).

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 find in Figure 3(b) that more intense flagging generated no significant increase in resources budgeted to flagged schools (RDA, $p > 0.2$), and only a minor increase in expenditures from teacher presence flagging (18%, $p = 0.09$, equivalent to roughly 75 US dollars per school-year; all other estimates $p > 0.1$). Nor do we find corresponding actions to improve service delivery through more visits or rotation of head teachers (RDA, mixed signs, all $p > 0.1$ except three with $0.05 < p < 0.1$; Supplementary Table C15), consistent with the null we document on final outcomes. See Supplementary Information C.4 for details and additional results.

Figure 3: Effect of flagging accountability in the command-and-control scheme



Notes: **(a)**, Coefficient and 95% confidence interval from estimating equation 3. The outcome is displayed in the y-axis, measured from 0 to 100 in year $y+1$. Flagging is based on the same outcome in year y. Blue points are for ordinary least squares (OLS) and gray for regression discontinuity aggregation (RDA), which captures the causal effect. The effects are interpreted as the percentage point change from a one-standard-deviation increase in flagging. The panel reports the number of observations (school-years), the first stage F-statistic, the mean and standard deviation of the number of times flagged in year y. **(b)**, Same design and sample sizes as **(a)**. Top row results are estimated on the share of yearly school visits and the share that had a change of head teacher. The bottom reports on the (log) school development funds and expenditure. Outcomes in **(a)** and **(b)** are measured in year $y+1$. The flagging variable is reported on the y-axis. **(c)**, Coefficient and 95% confidence interval for the event-study based on equation 4 using -1 as the base period, comparing schools in flagged and non-flagged *maraakiz*. The blue line uses the full sample. The red line uses the threshold sample, obtained through regression discontinuity methods. The results are for flagging on the variable in each plot. *Flag* is the period when the *markaz* is flagged. *Punish* is the period where flagging reports are distributed and punishment occurs. *After flag* is for periods after the oversight meeting occurs. All panels confidence intervals are based on standard errors clustered at the *markaz* level.

4 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 naively misread the data they collect. Education outcomes that fall sharply due to a temporary shock tend to recover on their own (Chay, McEwan and Urquiola, 2005). A senior manager who intervenes after such a shock and then observes the bounce-back may credit the intervention rather than the underlying mean reversion.

As described in Section 2, performance management was organized around flagging underperforming jurisdictions and reviewing their outcomes in subsequent periods through datapacks and regular accountability meetings. These datapacks showed current indicators alongside those from previous periods and targets, making performance in the most recent prior periods highly salient. Changes in performance following periods of underperformance were thus readily observable, and meeting discussions naturally focused on whether flagged jurisdictions had improved in subsequent periods. The system therefore lent itself to intuitive before–after comparisons which could lead post-flag improvements to be interpreted as evidence of program effectiveness.

Contemporaneous accounts suggest that such improvements were indeed interpreted in this way. Barber (2013), who advised the Chief Minister on the reform, notes that "The Secretary... [of the School Education Department], Aslam Kamboh, has firmly believed throughout that the rankings are the key to the progress we've made." (Barber, 2013, p. 45). Barber continues, "we introduced the "heat maps" which showed the Chief Minister which districts were on track and which weren't. ... The Chief Minister loved the heat maps: "I shall sleep with these under my pillow," he said." (Barber, 2013, p. 62).

Qualitative evidence further indicates the pervasiveness of belief in the system (Malik and Bari, 2023). One District Coordination Officer reflected: "Now I do not have qualitative or quantitative data so I can only tell you my impression. My impression is that attendance improved; teachers started coming to school...school facilities really improved" (Malik and Bari, 2023, p. 43). Another District Coordination Officer reported: "...they would show with green and red colors what was on track and what was not...there used to be a lot of accountability...that was very effective." (Malik and Bari, 2023, p. 32).

Our school-by-month panel lets us test this logic directly through a stacked difference-in-differences analysis. For every monthly *markaz* flagging event, we construct a sub-dataset of schools in the flagged *markaz* together with schools never flagged inside the same time window, and stack these

event panels in relative time around the flag (see Methods Section 7.3). Using the regression discontinuity logic from the previous section, we also construct a sample where we limit attention only to those *maraakiz* that fall within a 'close' bandwidth of the flagging threshold, with the idea that these *maraakiz* are more alike than the full sample, so should serve as better counterfactuals for each other (Calonico, Cattaneo and Farrell, 2020).

Figure 3(c) plots event study coefficients for these two samples, labeled 'full' and 'threshold' respectively, for each of the six targeted outcomes, with vertical lines marking the flag, the subsequent punishment meeting window, and the post-punishment period (see Supplementary Information D.1 for robustness). Focusing first on the full sample, we observe that flagged and non-flagged *maraakiz* follow similar paths just before the flagging. In the month of flagging, a shock reduces the outcome in the average school of flagged *markaz*, contributing to the *markaz* being selected for increased command and control accountability. However, as can be seen in the plots, by the punishment period these *maraakiz* have already mostly recovered to their historical trends. The shock is even more muted for the threshold sample, where the non-flagged *maraakiz* are arguably better counterfactuals. There is also no strengthening of recovery trends post-punishment for any outcome except teacher presence (one sided tests, $p > 0.1$; teacher presence $p < 0.01$, see Supplementary Information D.2 for a formal test). Student attendance and english scores have a slower recovery to pre-flag means compared to the rest of outcomes, which would be an important area of investigation for future work.

Viewed in isolation, flagged *maraakiz* trace exactly the arc the scheme was designed to produce: a sharp dip at the moment of flagging, oversight, and a visible recovery in the months that follow. These patterns are consistent with a senior manager comparing the flagging period with the post-punishment period and reading the increase as positive impacts of the scheme. Similarly, flagging may not have engendered a response from junior managers following the punishment period because the negatively shocked outcomes would already be returning to their historic trends. As we document, set against the non-flagged counterfactual *maraakiz*, the same recovery appears whether or not a *markaz* was ever flagged, with the "response" being reversion to the mean, not a product of command and control accountability.

An alternative interpretation is that the scheme may have sustained bureaucratic homeostasis, applying enough pressure to prevent deterioration without generating meaningful improvement. However, if homeostasis were driving the results, flagged *maraakiz* would recover differently from comparable non-flagged *maraakiz*. This interpretation is also less consistent with contemporaneous accounts across the hierarchy, which describe the scheme as producing improvement rather than merely preventing collapse.

5 Bayesian Interpretation of the Nulls

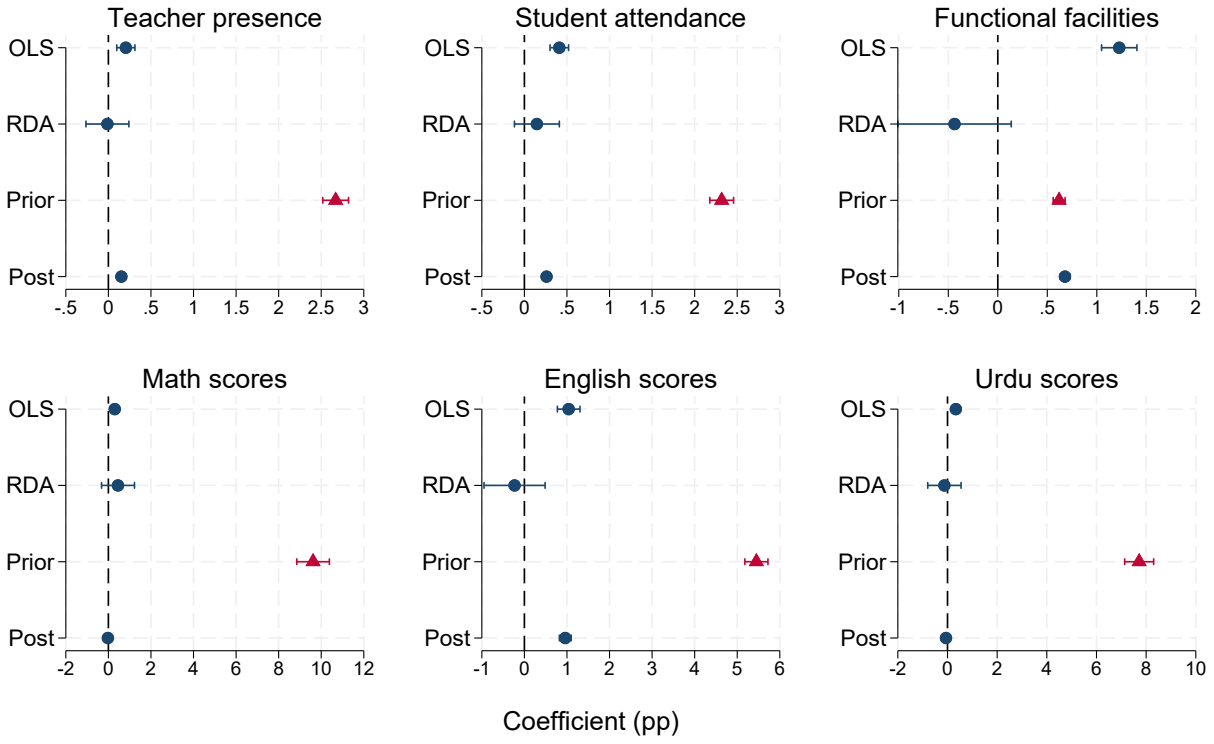
If senior managers misinterpret reversion to the mean as the positive effect of the scheme, how should our data analysis update their prior beliefs? We quantify this in a Bayesian setup by embedding the naive estimates described above - the before/after difference in outcomes for a flagged *markaz* — as a senior manager's prior about the effectiveness of flagging and update it with our causal estimates. See Methods Section 7.4 for details.

Figure 4 shows the result. The naive priors of senior managers place the scheme's effect several percentage points above zero (prior means of 2.6 to 9.6 percentage points, except functional facilities whose prior is 0.6 percentage points), consistent with the qualitative evidence we present in Section 4. But updating these beliefs with data that properly benchmarks flagged against non-flagged *maraakiz* yields a posterior close to zero (posterior means from -0.06 to 0.6 percentage points, with 95% credible intervals that include zero for every outcome), aligning almost entirely with our OLS and RDA estimates. Thus, a manager who mistook raw pre-post trends for program success would, after seeing the counterfactuals, revise that belief to essentially no effect.

6 Discussion

Centralized command of the public administration has been a dominant approach to the management of the public sector throughout history (Finer, 1997; Education Commission, 2023). Recent decades have seen its resurgence in a wave of data-driven, centralized monitoring and accountability schemes. U.S. examples span CompStat at the New York Police Department in 1994 and the Obama administration's 2009 Open Government Initiative, which produced a generation of federal dashboards (Behn, 2014; Ganapati, 2011). Similar centralized monitoring and accountability schemes have been adopted across the globe, including those of the United Kingdom (Bartlett and Tkacz, 2017), Ireland (Bartlett and Tkacz, 2017), the Netherlands (Matheus, Janssen and Maheshwari, 2020; van Rij, 2026), Brazil (Matheus, Janssen and Maheshwari, 2020), India (Duflo, Hanna and Ryan, 2012; Banerjee, Duflo and Glennerster, 2008), Pakistan (Callen et al., 2020), and Indonesia (Olken, 2007). Development institutions actively promote such systems: the World Bank's Government Analytics Handbook positions dashboards integrating administrative and survey data as a core tool for strengthening public administration (Rogger and Schuster, 2023). These dashboards, and the information systems that underlie them, have typically been used for centralized 'command and control' approaches to reforming government. Over 150 such initiatives have been documented across more than 80 countries, with adoption accelerating since 2010, nearly half focused on education, and the majority relying on external funding and technical assistance

Figure 4: Bayesian prior and posterior comparison



Note: This figure presents point estimates and 95% intervals from four estimators. Coefficients for ordinary least squares (OLS) and regression discontinuity aggregation (RDA) are estimated using equation 3, as reported in Figure 3**(a)** and Supplementary Table C1. The coefficient for senior managers' *Prior* is estimated using equation 4, for the effect of monthly flagging between the flag and punish periods as described in Supplementary Information D. It captures the naive belief in the effectiveness of flagging. *Post* is the posterior mean of the flagging effect from a Bayesian model that combines this prior with the data. We use a Gaussian likelihood for the outcomes. The prior on the flagging coefficient is normal, centered on the *Prior* estimate with variance equal to its squared standard error. We regularize and shrink *markaz* and year fixed effects towards their mean using a hierarchical normal prior centered at zero with 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 simulated by Markov chain Monte Carlo using the Metropolis–Hastings algorithm, with a single chain of 12,500 iterations of which the first 2,500 are discarded as burn-in. Error bars are 95% confidence intervals based on standard errors clustered at the *markaz* level for OLS, RDA and *Prior*, and equal-tailed 95% credible intervals for *Post*.

(Mansoor et al., 2021).

This wave of reform reflects a broader empirical regularity in public administration: substantive change rarely arrives through formal de jure rules, which face well-documented challenges of sustained adoption (Grindle, 2004; Peters, 2018; Pritchett, Woolcock and Andrews, 2013; Andrews, 2013; Banerjee et al., 2021; Muralidharan and Singh, 2020; Williams, 2026), but instead operates through de facto incentive structures (Lipsky, 1980; Schuster et al., 2023; Mangla, 2022). Despite limited evidence of their efficacy, these schemes remain attractive because they require no change to civil-service rules, they produce visible artifacts of action for political principals, and they are

easy for donors to fund and external advisors to deliver. Yet the data they produce also provide a novel opportunity to evaluate key components of such systems. This paper evaluates the accountability component of a flagship command and control scheme implemented in the education sector of Pakistan's Punjab Province, providing one of the first at-scale assessments of such reforms.

Our analysis reinforces the scholarly literature's concerns about the efficacy of such approaches. Punjab's system was designed to connect school-facing oversight by AEOs with district- and provincial-level authority over personnel, discipline, and resources, coordinated through recurring hierarchical review. Yet, the flagging-triggered activation of this hierarchy did not produce substantive impacts on school outcomes along any targeted indicator. Across the chain of service delivery, we find negligible impacts on government functioning beyond a slight increase in school operating expenditures. Our data allow us to make these claims with a high degree of precision. The results suggest that centralized command and control management approaches to accountability and control struggle to effectively manage unpredictable delivery environments.

Such findings echo dissenting sentiments expressed by members of the administration itself, as documented by Malik and Bari (2023). One sub-district officer involved in the scheme stated, "when these meetings used to happen, we used to forget the real purpose, and we would get busy trying to fulfill these tasks [that we were asked to do]. The real purpose is to teach the children, and that would get neglected. The tension of the meetings and monitoring was the least effective [aspect of the reform] because the entire work week would go into preparing for these meetings." (Malik and Bari, 2023, p. 41) Another stated, "The least effective were the limitless meetings … we would move away from the main point of focus and we would start focusing on unimportant things." (Malik and Bari, 2023, p. 41) As Malik and Bari (n.d.) themselves note, "It is evident though that [because of] the data seen during the stock-take meetings, actions were taken to course correct. Our interpretation is that these are firefighting responses and not necessarily sustained or ones that bring about structural improvements."

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, we find that all transitions after flagging reflect an organic reversion-to-the-mean.

Our evidence does not establish that mistaken before-after inference was the sole reason the system persisted; centralized monitoring may also serve internal political or organizational purposes. In

Punjab, however, contemporaneous accounts directly document a widespread belief that the system was improving education, and the monthly data show how that belief could arise. Moreover, using a Bayesian framework, we show that had the data feeding the system been used to update senior managers beliefs about the scheme, they would have rapidly observed the limitations of the accountability mechanism within the centralized command and control management.

7 Methods

7.1 Data

For the empirical analysis we use administrative data on public education collected by the Program Monitoring and Implementation Unit (PMIU). These data consist of monthly reports from roughly 50 thousand primary and middle schools between December 2011 and March 2018. June, July, and August of each year are excluded due to schools being closed, so we measure years as 'academic year' covering the nine months (September to May).

Data collection visits occur on random unannounced dates by Monitoring and Evaluation Assistants, who are staff from an different reporting line and whose primary job is facility inspections in health and education. The collected data includes the overall number of teachers and students, the number present on the day of the visit, indicators for the functionality of schooling infrastructure, the history of previous bureaucratic visits, and various school characteristics. Since 2016, the data also include measures of test scores, collected by randomly selecting seven Grade 3 students to take a standardized test on English, Math, and Urdu at the time of the data collection visit. The data are used to generate monthly reports that summarize key metrics based on which *maraakiz* are flagged or not on each performance indicator. We use the data to build our schooling outcome measures. The resulting school-by-month panel across more than three thousand *maraakiz* is unbalanced because of the creation of *maraakiz* over time.

Datapacks Our main outcomes are derived from monthly assessments of teacher presence, student attendance, and school facility functionality. The first two are measured as the share of teachers and students present during the unannounced data collection visit. Functional facilities measures the operational status of drinking water infrastructure, electricity, toilets, and boundary walls, aggregated as the share of facilities in working condition. Student test scores are measured as the share of correct answers in standardized tests in Math, English, and Urdu. Monitoring and Evaluation Assistants visited all *maraakiz* monthly, but school-level data were collected for 88% of all

possible school-month observations. We impute missing values using the nearest available month with non-missing data, leveraging the random visit allocation schedule. Results are robust to using the non-imputed data (Supplementary Information C.1, C.4, and D).

Performance thresholds School performance was flagged against centrally defined thresholds. Teacher presence was coded red below 86%, orange between 86% and 90%, and green at or above 90%. The corresponding thresholds for functional facilities were 90% and 95%, and for test scores 70% and 80%. Student attendance thresholds varied by district performance category and season. Districts are classified into three tiers: Category A comprises historically high-performing districts, Category C the lowest-performing, and Category B the remainder. The seasons are into high-attendance months (December to March) and low-attendance months (April to November). For Category A districts, attendance was coded red below 89% and green at or above 92% during December-March, and 87% and 90% during April-November. Category B districts faced thresholds of 87% and 90% in December-March and 85% and 88% in April-November. Category C thresholds were 84% and 87% in December-March and 82% and 85% in April–November.

Data compliance and quality We have datapack reports for all months from December 2011 through May 2018. The institutional separation between monitoring and implementation within the education administration ensured that reporting incentives did not distort measured outcomes. Data collection was further subject to monthly audits by district-level supervisors and quarterly third-party validation, lending additional credibility to the collected information (Khan, 2018).

We independently validate data quality by comparing the datapacks against the Annual School Census, the government's primary source of administrative statistics on the education sector. Using the information on posted teachers, enrolled students, and school infrastructure functionality, the two sources yield overlapping distribution. This confirms that the datapacks accurately describe the target population and further support its reliability for our empirical analysis. Further details are provided in Supplementary Information B.1.

7.2 Intensity of Centralized Accountability

We study how a more intense application of accountability in the command and control scheme in year y affects educational outcomes in year $y+1$. We measure accountability intensity by the number of months in year y in which a *markaz* is flagged red, which triggers the scrutiny, reprimands, and sanctions that constitute the accountability treatment in our setting. The challenge is that this

variable is not randomly assigned: chronically underperforming *maraakiz* are mechanically flagged more often, conflating the effects of accountability with underlying local conditions.

The RDA approach We address this endogeneity concern using the Regression Discontinuity Aggregation (RDA) design formalized by Borusyak and Kolerman-Shemer (2024). RDA extends standard regression discontinuity (RD) to settings in which the treatment is an aggregate across multiple discontinuity events. In a standard RD, a unit whose running variable falls barely above or below a threshold provides a single local experiment that, under a local randomization assumption (Cattaneo, Frandsen and Titiunik, 2015), generates as-if-random variation in treatment assignment. In our setting, each month provides one such experiment: a *markaz* is flagged or not depending on whether its aggregate performance crosses a fixed threshold. To measure how intensely the system is applied over a school year, we generate the treatment at the *markaz*-year level, summing across up to nine monthly flags. The RDA approach leverages this structure by instrumenting the number of times a *markaz* is flagged in a year with the monthly RD experiments as we describe next.

Estimation equation We estimate the following model:

$$Y_{s,m,d,y+1} = \beta \cdot \text{TimesFlagged}_{m,d,y} + \alpha_m + \lambda_y + \delta_{dy} + \varepsilon_{s,m,d,y+1} \tag{1}$$

where $Y_{s,m,d,y+1}$ is the average outcome for school s in *markaz* m and district d in school year $y+1$. $\text{TimesFlagged}_{m,d,y}$ is the standardized (z-score) number of months in year y in which the *markaz* is flagged, so that a one-unit increase represents a one-standard-deviation increase in flagging intensity over mean flagging. α_m are *markaz* fixed effects, λ_y are year fixed effects, δ_{dy} are district-specific linear time trends, and $\varepsilon_{s,m,d,y+1}$ is the error term, clustered at the *markaz* level.

To construct the instrument, we first identify, for each month and outcome, *maraakiz* whose aggregate performance lies within a small bandwidth on either side of the flagging threshold. We select the bandwidth using the standard RD procedure of Calonico, Cattaneo and Farrell (2020), computed separately for each outcome. For a given *markaz*-year, a month contributes to the instrument if the *markaz* was both close to the threshold *and* flagged (that is, fell just below the threshold). The instrument, $\text{CloseFlagged}_{m,d,y}$, is the share of months in year y in which the *markaz* was narrowly flagged. This resembles a shift-share where RD experiments are exogenous shifts and the share of each month in a year (equally weighted) are exposure shares (see Borusyak and Kolerman-Shemer (2024) for details).

Following Borusyak and Kolerman-Shemer (2024), we also include three RDA controls that aggre-

gate the standard local-linear RD covariates to the *markaz*-year level. First, $\text{Close}_{m,d,y}$ is the share of months in which the *markaz* was close to the threshold (on either side). This control ensures that the instrument isolates variation only among units with comparable exposure to close events. Second, $\text{Close}_{m,d,y} \times \bar{X}_{m,d,y}$, which is the share of months a *markaz* is in the threshold window multiplied by the average running variable in those months. Third, $\text{CloseFlagged}_{m,d,y} \times \bar{X}_{m,d,y}$: the share of months a *markaz* is in the threshold window and flagged multiplied by the average running variable in those months. The latter two are aggregates of the local-linear controls standard in RD estimation. They reduce bias from bandwidth selection, mirroring the role of the running variable and its interaction with treatment status in a conventional RD specification.

We estimate the following first stage via two-stage least squares:

$$\begin{aligned}\text{TimesFlagged}_{m,d,y} =& \gamma \cdot \text{CloseFlagged}_{m,d,y} + \gamma_1 \cdot \text{Close}_{m,d,y} + \gamma_2 \cdot (\text{Close}_{m,d,y} \times \bar{X}_{m,d,y}) \\ &+ \gamma_3 \cdot (\text{CloseFlagged}_{m,d,y} \times \bar{X}_{m,d,y}) + \alpha_m + \lambda_y + \delta_{dy} + u_{m,d,y} \end{aligned} \tag{2}$$

where $\text{CloseFlagged}_{m,d,y}$ is the excluded instrument, the three γ_1–γ_3 terms correspond to the RDA controls, and $u_{m,d,y}$ is the error term clustered at the *markaz* level. The second stage is:

$$\begin{aligned}Y_{s,m,d,y+1} =& \beta \cdot \widehat{\text{TimesFlagged}}_{m,d,y} + \beta_1 \cdot \text{Close}_{m,d,y} + \beta_2 \cdot (\text{Close}_{m,d,y} \times \bar{X}_{m,d,y}) \\ &+ \beta_3 \cdot (\text{CloseFlagged}_{m,d,y} \times \bar{X}_{m,d,y}) + \alpha_m + \lambda_y + \delta_{dy} + \varepsilon_{s,m,d,y+1} \end{aligned} \tag{3}$$

where $\widehat{\text{TimesFlagged}}_{m,d,y}$ is the fitted value from equation (2). Our parameter of interest is β, the causal effect of one standard deviation increase of flagging intensity, driven by months in which the *markaz* fell just below the performance threshold.

Identifying assumptions To estimate a causal effect we require the instrument to be 'as-if' randomly assigned. We study this assumption first by testing for the manipulation of flagging assignment using Cattaneo, Jansson and Ma (2020) discontinuity in the density approach. We find no evidence for manipulation. We then test the (quasi) random assignment assumption by examining if the treatment variable and instrument correlate with lagged $(y-1)$ outcomes conditional on the RDA controls and fixed effects from the main specification. The results show that the instrument

is not correlated with past performance, and thus plausibly orthogonal.

We also require the exclusion restriction assumption to hold. This assumption states that the share of months in the school-year that a *markaz* is flagged close to the threshold should influence subsequent school performance only through its effect on the number of times the *markaz* is flagged. This assumption is plausible because the thresholds were set arbitrarily, based on broad, district-wide performance distributions rather than pre-existing *markaz*-level cutoffs, and were therefore 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 instrument is unlikely to affect school performance through any channel other than flagging intensity.

Finally, we find that the instrument explains overall flagging and displays a positive and monotonic relationship with the endogenous variable, residualized from the fixed effects reported in equation 2, suggesting that the relevance and monotonicity assumption are met.

Further details of the identification assumptions results are available in Supplementary Information C.2.

RDA approach in other settings Our approach to identification is similar to prior literature estimating causal effects from multiple discontinuity events. For instance, research on political selection and representation have used within-party close-elections as RD shocks aggregated at the council or municipal level (Folke, 2014; Freier and Odendahl, 2015; Hyytinen et al., 2018; Meriläinen, 2022; Sørensen, 2023; Baskaran, Hessami and Schirner, 2024; Geys, Murdoch and Sørensen, 2024), and within-constituency close-elections aggregated at the district or state level (Clots-Figueras, 2011, 2012; Bhalotra and Clots-Figueras, 2014; Bhalotra et al., 2014; Nellis, Weaver and Rosenzweig, 2016; Nellis and Siddiqui, 2018; Priyanka, 2020; Bhalotra, Clots-Figueras and Iyer, 2021) to estimate effects on public sector performance. Similarly, research on growth and innovation has used travel distance discontinuities aggregated at the airport level as an instrument for position in global travel network (Campante and Yanagizawa-Drott, 2018; Bahar et al., 2023), and funding thresholds aggregated at the patent level to instrument overall funding access (Azoulay et al., 2019).

Across all studies, the common logic is that causal variation at aggregate levels can be obtained from many individual regression discontinuities. Borusyak and Kolerman-Shemer (2024) build on this logic to formalize the assumptions and requirements. In particular, the authors argue the RDA differs from prior literature in the extent to which they additionally account for aggregates of local-linear controls to closely mirror standard RD designs. Notably, they argue that the RDA can

be used to recover variation from RD events over time, as we do in our setting when aggregating monthly RD shocks at the year level.

Robustness Supplementary Information C.1 shows that our results are robust to a number of tests, including alternative instrument definitions, cross-flagging between outcomes and sample restrictions. We also explore effects by year and find no trends suggestive of flagging effect accumulating over time.

7.3 Naive Evaluations of Response

We study if the limited impact of flagging in the command and control scheme is driven by a natural return of the outcomes to their original state. To do so, we implement a stacked difference-in-discontinuities analysis, where we compare monthly outcomes for schools in flagged and non-flagged *maraakiz*, before and after the month when a flag occurs. The difference-in-discontinuities allow us to compare *maraakiz* in the threshold sample, defined as those *maraakiz* that, at the time of flagging, were narrowly above and below the threshold. We also report results for the full sample. The stacking avoid biases driven by the time-varying nature of the treatment (de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021) and allows us to estimate the immediate dynamic responses to flagging.

We estimate the following event study equation to examine the extent to which the null effects are driven by a return of outcomes to their original levels:

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

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. β_j captures the effect of being in a flagged *markaz* for each relative time $t = j$. α_{me} is for *markaz* fixed effects, and λ_{te} is for time fixed effects. We also include δ_{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 and present robustness for alternative stacking and specifications in Supplementary Information D.1.

Identifying assumptions Identification relies on no anticipation and parallel trends. The no anticipation assumption requires schooling outcomes not to respond to the expectation of the *markaz*

being flagged. The event studies in Figure 3(c) show no significant or small in magnitude pre-coefficients, suggesting there was no anticipation to the flagging. In addition, the facts that the negative shock leading to flagging is temporary and unexpected, and the punishment itself does not improve outcomes beyond the pre-flag mean, suggest that any anticipatory behavior is negligible. The parallel trends require that school outcomes in flagged and not-flagged *maraakiz* would have followed the same path in the absence of the command and control scheme. While this assumption is not directly testable, the pre-coefficients for the full and threshold sample in Figure 3(c) suggest it is plausible. Even more, the threshold sample addresses possible violations of the parallel trends as schools in *maraakiz* close to the threshold are more comparable, ensuring that they are similar in characteristics except for the (quasi) random flagging status.

Robustness In Supplementary Information D.1 we show that our results are robust to a host of tests. We test if our results are sensitive to anticipation of the command and control scheme implementation; to variations in the stacked-data structure; to responses around the orange/green flagging thresholds and defining flagging at higher administrative units. We also test for path dependence and prior exposure to flagging; alternative difference-in-differences estimators that explicitly account for staggered treatment timing and treatment switching on and off; and a standard regression discontinuity specification. Finally, we examine if the return to pre-treatment means is some combination of mean reversion and the impact of the punishment period using first-differenced outcomes.

7.4 Bayesian Interpretation of the Nulls

To quantify how the data would have updated a senior managers prior beliefs about the flagging effectiveness 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. We use equation 4 to compare the flagging and punishment periods, between which the reversion to the mean occurs. 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. The posterior distributions (effects updated by the priors) are calculated through Markov Chain Monte Carlo simulations using the Metropolis-Hastings algorithm.

We treat fixed effects as nuisance parameters and regularized then 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 overfit-

ting and noisy estimates (Athey and Imbens, 2017). Our sample size makes this unlikely, but the hierarchical regularization improves precision due to the high-dimensionality of our fixed effects.

References

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, Matt. 2013. *The Limits of Institutional Reform in Development: Changing Rules for Realistic Solutions.* Cambridge:Cambridge University Press.

Argyris, Chris, and Donald A. Schön. 1997. "Organizational Learning: A Theory of Action Perspective." *Reis*, , (77/78): 345–348.

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, S., and G.W. Imbens. 2017. "Chapter 3 - The Econometrics of Randomized Experimentsa." In *Handbook of Field Experiments*. Vol. 1 of *Handbook of Economic Field Experiments*, , ed. Abhijit Vinayak Banerjee and Esther Duflo, 73–140. North-Holland.

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.

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." Policy report, accessed May 18, 2026.

Bartlett, Jamie, and Nathaniel Tkacz. 2017. "Governance by Dashboard: A Policy Paper." Demos, London. accessed May 18, 2026.

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

Bechky, Beth A., and Daisy E. Chung. 2018. "Latitude or Latent Control? How Occupational Embeddedness and Control Shape Emergent Coordination." *Administrative Science Quarterly*, 63(3): 607–636.

Behn, Robert D. 2003. "Why Measure Performance? Different Purposes Require Different Measures." *Public Administration Review*, 63(5): 586–606.

Behn, Robert D. 2014. *The PerformanceStat Potential: A Leadership Strategy for Producing Results*. Washington, DC:Brookings Institution Press. accessed May 18, 2026.

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

Bhalotra, Sonia, and Irma Clots-Figueras. 2014. "Health and the Political Agency of Women." *American Economic Journal: Economic Policy*, 6(2): 164–97.

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. Economics, Religion, and Culture.

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.

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

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(Volume 26, 2023): 271–290.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230. Themed Issue: Treatment Effect 1.

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., Brigham R. Frandsen, and Rocio Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. 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.

Chaudhry, Rastee, and Abdullah Waqar Tajwar. 2020. "The Punjab Schools Reform Roadmap: A Medium-Term Evaluation." *Implementing Deeper Learning and 21st Education Reforms*, 109–128. 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.

Cilliers, Jacobus, and James Habyarimana. 2023. "Tackling Implementation Challenges with Information: Experimental Evidence from a School Governance Reform in Tanzania." RISE RISE Working Paper 23/142. accessed May 18, 2026.

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

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'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 2964–96.

Deci, Edward L., and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior.* Springer US.

Deserranno, Erika, A. Stefano Caria, Philipp Kastrau, and Gianmarco León-Ciliotta. 2025. "The Allocation of Incentives in Multilayered Organizations: Evidence from a Community Health Program in Sierra Leone." *Journal of Political Economy*, 133(8): 2506–2562.

Deserranno, Erika, Philipp Kastrau, and Gianmarco León-Ciliotta. 2025. "Promotions and Productivity: The Role of Meritocracy and Pay Progression in the Public Sector." *American Economic Review: Insights*, 7(1): 71–89.

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.

Dodge, Eric, Yusuf Neggers, Rohini Pande, and Charity M Troyer Moore. 2025. "From Delay to PayDay: Easing Bureaucrat Access to Implementation Information Strengthens Social Protection Delivery." National Bureau of Economic Research Working Paper 33756.

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

Education Commission. 2023. "Deliberate Disrupters: Can Delivery Approaches Deliver Better Education Outcomes?" Education Commission DeliverEd Final Report. accessed May 18, 2026.

Espeland, Wendy Nelson, and Michael Sauder. 2007. "Rankings and Reactivity: How Public Measures Recreate Social Worlds1." *American Journal of Sociology*, 113(1): 1–40.

Eyben, Rosalind. 2010. "Hiding Relations: The Irony of 'Effective Aid'." *The European Journal of Development Research*, 22: 382–397.

Falk, Armin, and Michael Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review*, 96(5): 1611–1630.

Finan, F., B.A. Olken, and R. Pande. 2017. "Chapter 6 - The Personnel Economics of the Developing State." In *Handbook of Economic Field Experiments*. Vol. 2 of *Handbook of Economic Field Experiments*, , ed. Abhijit Vinayak Banerjee and Esther Duflo, 467–514. North-Holland.

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.

Freidson, Eliot. 2001. *Professionalism, the third logic: On the practice of knowledge*. University of Chicago press.

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.

Ganapati, Sukumar. 2011. "Use of Dashboards in Government." IBM Center for The Business of Government. accessed May 18, 2026.

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. Themed Issue: Treatment Effect 1.

Grindle, Merilee S. 2004. "Good Enough Governance: Poverty Reduction and Reform in Developing Countries." *Governance*, 17(4): 525–548.

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.

Honig, Dan. 2024. *Mission Driven Bureaucrats: Empowering People To Help Government Do Better.* Oxford University Press.

Hood, Christopher. 1991. "A PUBLIC MANAGEMENT FOR ALL SEASONS?" *Public Administration*, 69(1): 3–19.

Hood, Christopher. 2006. "Gaming in Targetworld: The Targets Approach to Managing British Public Services." *Public Administration Review*, 66(4): 515–521.

Hussain, Iftikhar. 2015. "Subjective Performance Evaluation in the Public Sector." *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.

Kerr, Steven. 1995. "On the Folly of Rewarding A, while Hoping for B." *The Academy of Management Executive (1993-2005)*, 9(1): 7–14.

Kettl, Donald F. 2005. *The Global Public Management Revolution.* . 2 ed., Washington, DC:Brookings Institution Press.

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, Kashmali. 2018. "Punjab, Pakistan: Using Open School Data to Improve Transparency and Accountability." UNESCO International Institute for Educational Planning (IIEP), Paris, France. Open access publication.

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.

Lipsky, Michael. 1980. *Street Level Bureaucracy: Dilemmas of the Individual in Public Services.* Russell Sage Foundation.

Malik, Rabea, and Faisal Bari. 2023. "Improving Service Delivery Via Top-Down Data-Driven Accountability: Reform Enactment of the Education Road Map in Pakistan." Education Commission and Blavatnik School of Government DeliverEd Initiative Working Paper. accessed May 18, 2026.

Malik, Rabea, and Faisal Bari. n.d.. "Qualitative source materials for Improving Service Delivery Via Top-Down Data-Driven Accountability: Reform Enactment of the Education Road Map in Pakistan." Unpublished interview excerpts and research notes underlying Malik and Bari (2023), shared privately with the authors.

Mangla, Akshay. 2022. *Making Bureaucracy Work: Norms, Education and Public Service Delivery in Rural India. Cambridge Studies in the Comparative Politics of Education*, Cambridge University Press.

Mansbridge, Jane. 2014. "A Contingency Theory of Accountability." In *The Oxford Handbook of Public Accountability*. , ed. Mark Bovens, Robert Goodin and Thomas Schillemans. Oxford University Press.

Mansoor, Zahra, Dana Qarout, Kate Anderson, Celeste Carano, Liah Yecalo-Tecle, Veronika Dvorakova, and Martin J. Williams. 2021. "A Global Mapping of Delivery Approaches." Education Commission and Blavatnik School of Government DeliverEd Initiative Working Paper. accessed May 18, 2026.

Matheus, Ricardo, Marijn Janssen, and Devender Maheshwari. 2020. "Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities." *Government Information Quarterly*, 37(3): 101284.

McCubbins, Mathew D., Roger G. Noll, and Barry R. Weingast. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics, & Organization*, 3(2): 243–277.

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

Mintzberg, Henry. 1979. "The structuring of organizations: A synthesis of the research." *Englewood Cliffs, NJ*.

Muller, Jerry Z. 2018. *The Tyranny of Metrics.* Princeton University Press.

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

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, and Steven C. Rosenzweig. 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.

Pakistan Bureau of Statistics. 2023. "7th Population and Housing Census-2023: Provincial Census Report – Punjab." Pakistan Bureau of Statistics, Government of Pakistan Technical Report, Islamabad, Pakistan. accessed April 27, 2026.

Peters, B. Guy. 2018. *The Politics of Bureaucracy: An Introduction to Comparative Public Administration.* . 7th ed., New York:Routledge. accessed May 18, 2026.

PMIU-PESRP. 2025. "Annual School Census 2024–25: Punjab's Public Education: Key Trends & Insights." Programme Monitoring and Implementation Unit (PMIU-PESRP), School Education Department, Government of the Punjab, Lahore, Pakistan. Accessed: 2026-04-27.

Pollitt, Christopher, and Geert Bouckaert. 2017. *Public Management Reform: A Comparative Analysis—Into the Age of Austerity.* . 4th ed., Oxford:Oxford University Press. accessed May 18, 2026.

Power, Michael. 1999. *The Audit Society: Rituals of Verification.* Oxford University Press.

Prendergast, Canice. 2002. "The Tenuous Trade-off between Risk and Incentives." *Journal of Political Economy*, 110(5): 1071–1102.

Pritchett, Lant, Michael Woolcock, and Matt Andrews. 2013. "Looking Like a State: Techniques of Persistent Failure in State Capability for Implementation." *The Journal of Development Studies*, 49(1): 1–18.

Priyanka, Sadia. 2020. "Do female politicians matter for female labor market outcomes? Evidence from state legislative elections in India." *Labour Economics*, 64: 101822. European Association of Labour Economists, 31st annual conference, Uppsala Sweden, 19-21 September 2019.

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. 2021. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory*, 31(2): 259–277.

Rogger, Daniel, and Christian Schuster. 2023. *The Government Analytics Handbook: Leveraging Data to Strengthen Public Administration.* The World Bank.

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.

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* New Haven, CT:Yale University Press. accessed May 18, 2026.

Simon, Herbert A. 1997. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations.* . 4 ed., New York:Free Press.

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

van Rij, Wybren. 2026. "Dashboards as boundary objects: Shaping data-policy boundaries in complex policy domains." *Government Information Quarterly*, 43(2): 102144.

Williams, Martin J. 2026. *Reform as Process: Implementing Change in Public Bureaucracies.* New York:Columbia University Press.

Wilson, James Q. 1989. *Bureaucracy: What Government Agencies Do and why They Do it.* Basic Books.

SUPPLEMENTARY INFORMATION

— For Online Publication —

Contents

A Centralized Oversight in Punjab **2**

A.1 Broad Context of Education Outcomes in Pakistan 2

A.2 The 'Delivery Approach' to Education Reform 3

A.3 Implementation Fidelity of the Oversight Scheme 5

A.4 The Scheme Had An Aura of Success Among Leading Architects 6

B Data **7**

B.1 Additional Information 7

B.2 Descriptive Statistics 8

C Intensity of Command-and-Control Accountability **13**

C.1 Additional Results 13

C.2 Identifying Assumptions 22

C.3 Alternative Margins of School Response 26

C.4 Changes in the Wider Machinery of Government 34

C.4.1 Bureaucratic Oversight Visits, Transfers and Postings 34

C.4.2 Budget Allocation and Utilization 35

D Naive Evaluations of Response **44**

D.1 Additional Results 44

D.2 Efficacy of Punishment 46

A Centralized Oversight in Punjab

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

As argued in the main paper, the command-and-control intervention was a major initiative of the Punjab Government between 2011 and 2018. The aim was to accelerate improvements in service delivery using regular data-driven monitoring routines linked with high-stakes bureaucratic accountability across the chain of hierarchy. This section provides greater detail on the reforms and elaborates qualitative details of the context, political support and implementation fidelity.

A.1 Broad Context of Education Outcomes in Pakistan

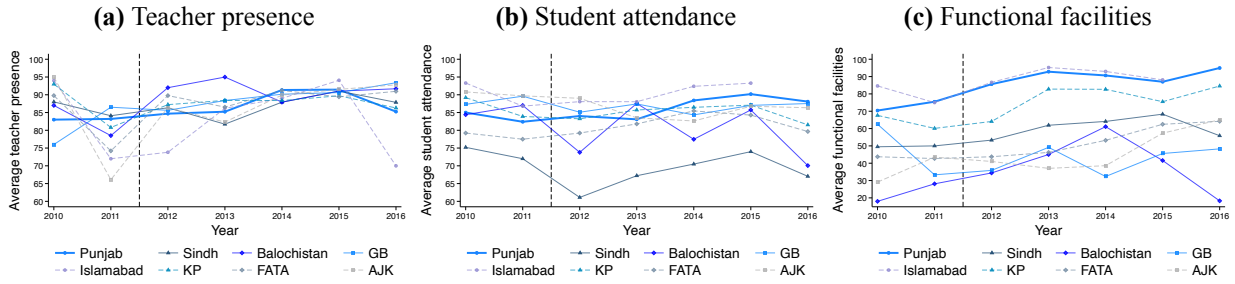
The broader dynamics of education in Pakistan over the period of interest were relatively stagnant. This can be seen in province-level data for the period 2010-2016 from the Annual Status of Education Report - ASER - Pakistan (`aserpakistan.org`). The ASER team have conducted independent and consistent household and school surveys to assess education progress in the country. Figure A1 shows the average trends in educational outcomes in all Pakistan provinces using ASER data in terms of teacher presence, student attendance and functional facilities. It can be observed that over the period of interest, there is at best a slight improvement across the country.

The presentation of comparative data provides an opportunity to reflect on how the scheme changed the dynamics of education in Punjab versus its neighbors. Similar schemes were not implemented in other provinces during the same period of time. Thus, at a descriptive level, if the command-and-control intervention had an impact on the overall trajectory of education outcomes in Punjab, it should be seen in its broad education dynamics relative to those in other provinces over the relevant time period. Such a trajectory does not seem to be distinguishable from Figure A1. Most provinces are improving or trending in a similar way to Punjab (darker blue line) over the period of interest.

To complement the cross-province comparison, we turn to monthly data for Punjab before and after the program's introduction to conduct a descriptive check for structural breaks at the scheme introduction. Consistent with the province-level comparison, Figure A2 does not point to a broad acceleration in outcomes at the scheme introduction. Teacher presence and student attendance shows no statistically significant break in level or trend. Functional facilities show no significant

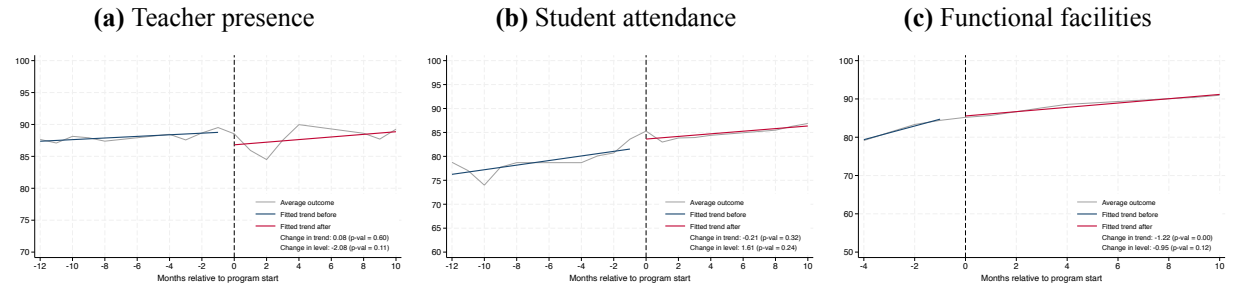
level break and but a trend break suggesting a slower evolution of outcomes. These descriptive patterns are consistent with our emphasis that the paper studies variation in accountability intensity within an already-operating scheme, rather than the effect of the scheme's introduction per se.

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. Punjab is represented by a darker blue solid line with other provinces represented by other lines. 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.

Figure A2: Punjab average monthly outcomes trends



Note: This figure presents a descriptive structural-break exercise around the introduction of the command-and-control scheme using monthly PMIU data for Punjab aggregated from the school-by-month level to monthly means. 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. Functional facilities pre-adoption data changed structure so we can only include up until relative time -4. Months are measured relative to the adoption month, marked by the dashed vertical line. The gray line plots observed monthly averages. The blue and red lines plot fitted linear trends before and after adoption. The change in trend at adoption is the coefficient on the interaction between months relative to adoption and a post-adoption indicator. The reported change in level at adoption is the coefficient on the post-adoption indicator. P-values are from the corresponding linear specification estimated on the monthly series.

A.2 The 'Delivery Approach' to Education Reform

The Delivery Approach was introduced in December 2010 during the political tenure of the Punjab Muslim League government under the helm of then Chief Minister, Shahbaz Sharif. As Malik and

Bari (2023) state, "The design for the approach drew on the theory and experience of delivery units globally, particularly Britain's experience." They continue, "In Punjab, the reform was structured to leverage the authority of the Chief Minister to create ownership of reform throughout the service delivery chain, and to catalyze large scale improvement in schooling access and quality over a relatively short period of time. ... This political oversight is from the top political leadership, very centralized; contrary to the role of politics in service delivery imagined in other systems levels models which view local politics and engagement with grassroots providers as central to design of accountability based reforms." (Malik and Bari, 2023, p. 23).

Priorities and targets for the Delivery Approach were set centrally. As Malik and Bari (2023) state, the targets did not take into account the specific "socio-political, geographical context of the district as well as [its] historical development trajectory" (Malik and Bari, 2023, p. 28). As such, quoting the District Coordination Officer for Mianwali, they state, "to fit all this [diversity] in a central system, a uniform way, a uniform criteria, was probably not possible."[a] Similarly, the District Coordination Officer for Chakwal was reported to have stated, "Perhaps in the provincial headquarters it seemed like the indicators were alright, but when you're in the field you realize there is a one jacket fit for all approach. ... One thing is that indicators should match ground reality." (Malik and Bari, 2023, p. 29)

The reform had the highest levels of political support. This is particularly true in the application of the monitoring of target achievement and accountability if targets weren't met. Education Commission (2023) writes 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" (Education Commission, 2023, p. 16). 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). A corresponding quote from the District Coordination Officer of Bahawalnagar states, "[The Chief Minister] might have many limitations but he also had a vast experience in working with different secretaries and bureaucrats. ... He used to hold all officers accountable. He used to stay very vigilant. He meant business!" And similarly, the district commissioner of Rawalpindi was reported as stating, "A major element of the approach was accountability. The Chief Minister used to be present in the meetings along with the chief secretary. So the officers would be very scared. ... public shaming used to happen."[b]

Overall, senior managers did not change de jure conditions to achieve these targets, such as making AEO salaries conditional on performance.[c] Instead, the system relied primarily on de facto

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

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

[c] Though some occasional ad hoc financial bonuses were given to district officials. We explore whether there is

incentives — long the dominant mode of motivation in public administration — with the aim of transmitting authority and discipline down the hierarchy.

A.3 Implementation Fidelity of the Oversight Scheme

The command-and-control scheme was implemented with high fidelity. This was underpinned by strong commitment by senior leadership and an effective rollout across the province. This section outlines qualitative evidence on those features of the scheme.

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).[d] 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). As noted earlier, 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

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.

[d] 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."

on the grass root level were the District Education Officers and the Assistant Education Officers ... These are officers who are in touch with schools at the markaz and individual school level."[e]

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."[fg]

A.4 The Scheme Had An Aura of Success Among Leading Architects

The Delivery Approach in Punjab has been highlighted by its architects 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." (Barber, 2017). He writes that such a perception was common within the leadership: "The Secretary... [of the School Education Department], Aslam Kamboh, has firmly believed throughout that the rankings are the key to the progress we've made." (Barber, 2013, p. 45); "The Chief Minister loved the heat maps: 'I shall sleep with these under my pillow,' he said" (Barber, 2013, p. 62); and, "The Chief Minister of Punjab, who admittedly has as powerful an interest as any in this work, called it a 'roaring success story'." (Barber, 2013, p. 26). Barber (2021) quotes the Secretary of School and Education Department again as having said, "At the start I was the biggest sceptic, in the end I was the greatest advocate." (Barber, 2021, p. 157)

Malik and Bari (2023) also find evidence that district-level officials believed the Delivery Approach had impacts too. The District Coordination Officer for Chakwal stated, "Now I do not have qualitative or quantitative data so I can only tell you my impression. My impression is that attendance improved; teachers started coming to school … school facilities really improved." (Malik and Bari, 2023, p. 43). He goes on, "an entire machinery was coming into sync under these education authorities. I believe that all of Punjab's results improved dramatically." (Malik and Bari, 2023, p. 32). The District Coordination Officer for Rawalpindi stated, "the ranking system was effective in some way. There was a sense of constant supervision [of the DC's responsibilities and the status of ser-

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

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

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

vice delivery]. ... So I think my simple answer to this would be that I think it did have an effect on the overall improvements that were brought into the system. ... performance accountability at regular intervals did help."[h] Another District Coordination Officer reported: "...they would show with green and red colors what was on track and what was not...there used to be a lot of accountability...that was very effective." (Malik and Bari, 2023, p. 32). And one district education officer believed that, "The most effective [thing] was that we improved teacher and student attendance." (Malik and Bari, 2023, p. 32)

B Data

We use administrative data on public education collected by the Program Monitoring and Implementation Unit (PMIU). These data consist of monthly reports from about 52,000 primary and middle schools between December 2011 and March 2018. Data collection visits occur on random unannounced dates by Monitoring and Evaluation Assistants, whose primary job is facility inspections in health and education. MEA reporting was subject to monthly audits by district-level supervisors and quarterly third-party validation, resulting in a high degree of data credibility (Khan, 2018). In this section we provide additional details, data quality assessments, and present descriptive statistics on the relevant school outcomes and flagging patterns.

B.1 Additional Information

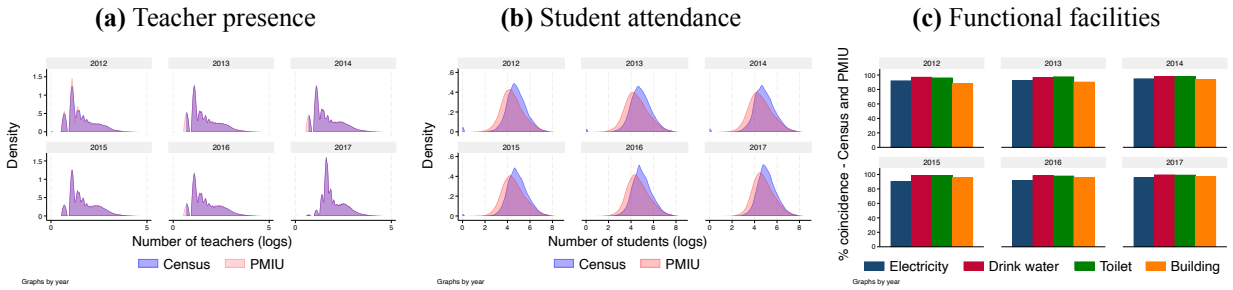
Our primary outcomes are generated from monthly assessments of teacher presence, student attendance, and whether school facilities are functional. The first two are measured as the percentage of teachers/students present at the time of the data collection visit. The functional facilities record the status of the drinking water infrastructure, electricity, toilets, and the boundary wall. We use the share of functional facilities as outcome. Scores are measured as the percentage of correct answers in Math, English and Urdu.

Data compliance and quality We assess the quality of the data-packs by comparing it against the census in October, when the census information was collected. Since data are not collected on the same day, nor by the same precise methods, there can be some measurement error. Figure B1 compares the distribution of the variables reported by both sources. Panels (a) and (b) show for teacher presence and student attendance that the data sources overlap almost completely. Panel (c) plots the percentage of schools where the functional infrastructure coincides, which is near 100%

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

for all the indicators. As such, there is no evidence of systematic manipulation, suggesting that the data-packs mapped consistently the population, supporting its reliability for the analysis.

Figure B1: Data validation - monthly PMIU vs. Census



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).

B.2 Descriptive Statistics

Table B1 shows statistics at the markaz-by-year level. 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%. Table B2 there is substantial variation in the number of schools within a markaz: average number of schools an AEO must manage is 22, of which 80% are elementary schools.

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 B2 shows the persistence of underperformance. Districts were ranked every quarter based on their overall performance, so the best (worse) 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. Panel C of Table B2 show statistics for the change in the ranking, further supporting the high persistence of the positions.

The systematic underperformance of some maraakiz aligns with evidence from other settings indicating that education face structural constraints (Bank, 2017). However, they are also exposed to shocks (such as teachers getting sick) that substantially shift the levels of service delivery. This would imply that Punjab's schools face shocks that might push maraakiz under the flagging thresh-

Table B1: Descriptive statistics - markaz-year

Panel A: Schooling outcomes (0-100)	Mean	Std. Dev.	Min.	Max.	Obs.
Teacher presence	93.095	4.778	0	100	15,035
Student attendance	90.030	5.739	0	100	15,037
Functional facilities	95.312	7.885	0	100	15,025
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	0.644	1.232	0	9	15,035
Student attendance - # times flagged	1.033	1.846	0	9	15,037
Functional facilities - # times flagged	1.009	2.255	0	9	15,025
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 as 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.

old irrespective of their baseline performance. The variation between schools is consistent with this interpretation. Table B3 presents the standard deviations in school outcomes by quintile of baseline performance. The top four quintiles face comparable levels of variation. On the contrary, variation in lowest quintile is larger, implying a persistently challenging environment to manage.

Instrument variation Figure B3 displaying the distribution of the number of times an school is in a flagged (blue bins) and close to the threshold flagged markaz (red bins). Both show that flagging was a relatively common feature of the education system and most schools were exposed to it.[i] Table B4 show descriptive statistics at the markaz-year level for each outcome annual exposure to flagging, and annual exposure to flagging while close to the threshold. The mean exposure to flagging ranges from 0.7 to 1.06 flags. Respectively, flagging around the threshold for the limits are 0.21 and 0.0.43, accounting for roughly 30 to 33 percent of the total flagging in a year.

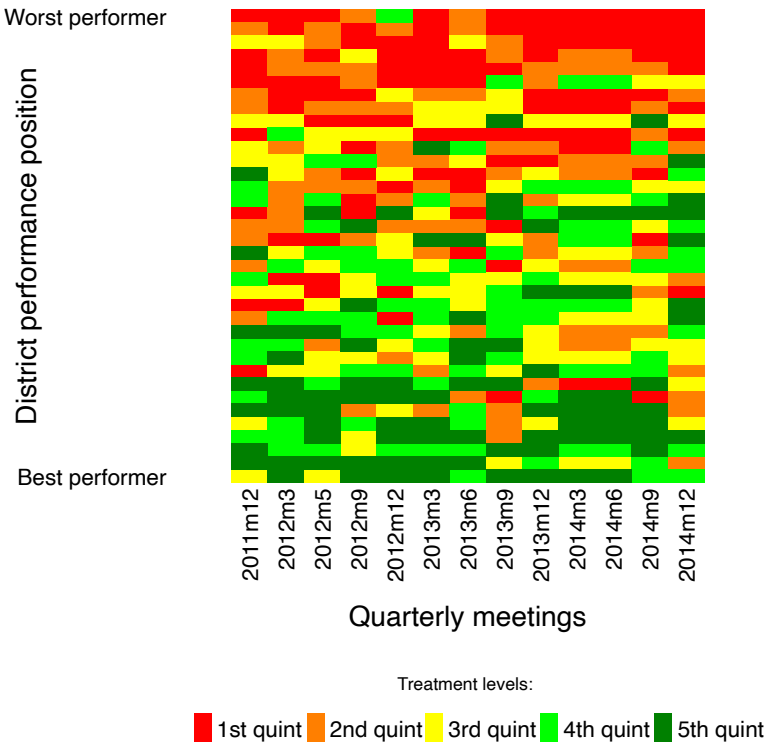
[i]For teacher presence, student attendance and functional facilities, a school can be in a flagged markaz at most nine times in an academic year. Between 2012 and 2018, it can be exposed at most 63 times per outcome, or 189 times in total. The highest number of times flagged by Math, English, and Urdu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

Table B2: 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	104,639	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.

Figure B2: Distribution of quintiles of district performance



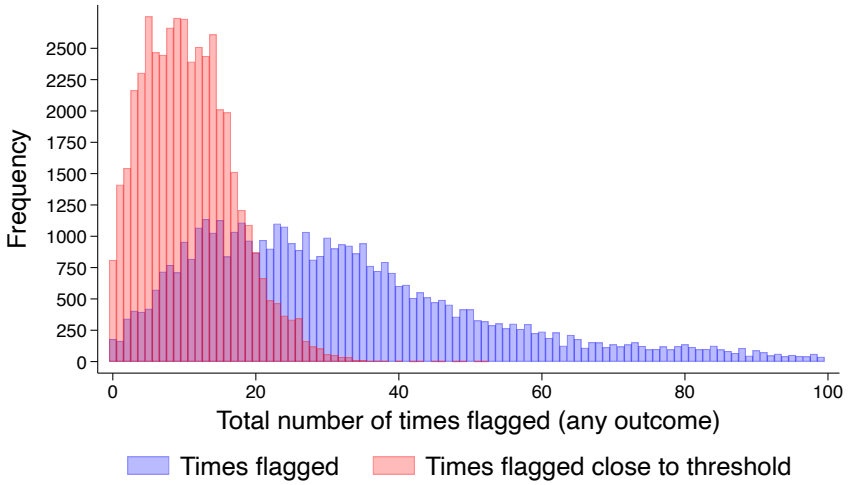
Note: This figure illustrates for each quarter the quintile of the overall district score distribution in which each district fell. District scores are measured based on the aggregate performance of teacher presence, student attendance, and functional facilities in each quarter. The figure ranks the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

Table B3: 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.

Figure B3: 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 B4 show detailed descriptive statistics at the markaz-by-year level for exposure to the flagging close to the threshold in each schooling outcome.

Table B4: Descriptive statistics - flagging in the threshold

Schooling outcome - flagging variable	Mean	Std. Dev.	Min.	Max.	Obs.
Teacher presence - # times close to the threshold	0.707	1.033	0	8	15,035
Teacher presence - # times flagged close to the threshold	0.212	0.561	0	7	15,035
Student attendance - # times close to the threshold	0.809	1.321	0	8	15,037
Student attendance - # times flagged close to the threshold	0.338	0.764	0	8	15,037
Functional facilities - # times close to the threshold	0.616	1.362	0	8	15,025
Functional facilities - # times flagged close to the threshold	0.218	0.711	0	8	15,025
Match score - # times close to the threshold	0.200	0.518	0	8	9,482
Match score - # times flagged close to the threshold	0.060	0.275	0	8	9,482
English score - # times close to the threshold	1.065	1.180	0	8	9,482
English score - # times flagged close to the threshold	0.435	0.715	0	5	9,482
Urdu score - # times close to the threshold	0.249	0.518	0	5	9,482
Urdu score - # times flagged close to the threshold	0.079	0.294	0	4	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.

C Intensity of Command-and-Control Accountability

We study the impact of flagging accountability in the command-and-control approach on public education performance in Section 3 of the main paper. The treatment is the intensity of exposure to flagging, measured as the count of months in the academic year in which a school was in a flagged markaz. To obtain a causal effect we rely on a regression discontinuity aggregation (RDA) design (Borusyak and Kolerman-Shemer, 2024) and instrument the treatment with the share of months in which flagging occurred close to the threshold. We define close to the threshold for each outcome independently using the minimum of the coverage-rate-optimal bandwidths Calonico, Cattaneo and Farrell (2020).[j] We estimate the effect of year y exposure on year $y+1$ outcomes. In this section, we present robustness and additional results.

C.1 Additional Results

Table C1 presents our main results, equivalent to Figure 3 (a). Panel A reports on school outcomes. Panel B on student scores. For each outcome at the top of the panel, the first column reports the OLS result and the second the RDA. Times Flagged$_y$ is defined as the z-score on each outcome flagging in year y, such that a unit increase is equal to a standard deviation increase in flagging intensity. We report the Kleibergen and Paap (2006) Wald test F-statistic (Andrews, Stock and Sun, 2019), which is high for all outcomes suggesting that the instrument is not weak.

OLS yields small and significant effects. For teacher presence, an increase of one standard deviation in flagging increases average teacher presence by 0.205 percentage points. Effects are similar and always smaller than a percentage point for all outcomes except functional facilities, which is close to a percentage point. The RDA coefficients indicate not significant improvements. Intuitively, teacher presence is the element of the production function that would be most responsive to hierarchical pressures. However, 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 small, and economically negligible, impacts.

[j] Cattaneo, Jansson and Ma (2020) argue that the commonly used MSE-optimal bandwidth is appropriate when the parameter of interest is a point estimate. However, because we are building a sample of RD experiments to aggregate, we are more interested on the coverage of confidence intervals, for which the CER-optimal bandwidth is more appropriate. The choice is conservative as it minimizes the bandwidth around the threshold for inference. The average bandwidth around the threshold is 2.65 pp for teacher presence, 2.73 pp for student attendance, 3.25 pp for functional facilities, 3.19 pp for Math, 4.31 pp for English, and 3.73 pp for Urdu.

Table C1: Impacts of exposure to flagging

Panel A: School outcomes						
	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.205***	-0.011	0.410***	0.147	1.227***	-0.437
	(0.054)	(0.129)	(0.055)	(0.134)	(0.091)	(0.292)
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		494.2		577.0		288.5
Panel B: Student scores						
	Dependent variables (range: 0-100)					
	Math$_{y+1}$		English$_{y+1}$		Urdu$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.301***	0.452	1.040***	-0.231	0.338***	-0.125
	(0.084)	(0.394)	(0.135)	(0.366)	(0.083)	(0.342)
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		39.5		220.3		122.4
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
RDA Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table presents the results from estimating equation 3 and is the table equivalent of Panel (a) Figure 3. Panel A reports on schooling outcomes. Panel B reports on student scores. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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 of the share of correct answers in standardized exams performed on a random sample of primary students. Standard errors clustered by markaz in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Alternative specifications Table C2 show the results with alternative definition of the instrument. Panel A uses the z-score of the number of times flagged close to the threshold instead of the share of months close to the threshold. Panel B uses nine dummy variables, one for each month of the school year, indicating if a markaz was flagged while being close to the threshold. This definition allow us to test for a non-parametric version of the instrument. In both cases we control for a time-varying fixed effect for the number of times close to the threshold instead of the RDA controls. This approaches are closer to the literature before Borusyak and Kolerman-Shemer (2024). Coefficients are no significant or lower than a percentage point, suggesting that our results are robust, further reinforcing the negligible effects of flagging in the command-and-control accountability. Table C3 uses as instrument indicators for ever flagged close to the threshold and flagged above the median intensity close to the threshold. Results are always robust.

Non-balanced school panel As discussed in Appendix B.1, most schools are visited every month. For the small percentage of school-month not receiving data collection visits we rely on the random visit allocation to impute the missing information with the closest month with non-missing data before aggregating the panel at the school-by-year level. Table C4 shows the results when aggregating the non-imputed data as a robustness. Compared to Table C1 results are very similar. Taken together the results are robust and the conclusions hold.

Other robustness First, we test if there are responses of flagging in one performance measure on a different schooling outcomes in Table C6. Most of the responses to cross-flagging are non-significant and smaller than a percentage point. Second, we test how the effects evolve by year in Figure C1. There are no clear time trends: teacher presence and student attendance remain flat and near zero over time, while functional facilities fluctuates between small positive and negative values. Third, we show in Table C5 that our results are robust to subsetting to the same sample as the orthogonality analysis in Table C7. Last, we find null effects when we vary bandwidth sizes (results on request).

Table C2: Impacts of exposure to flagging - alternative instrument definitions

Panel A: Instrument = # Times flagged close to threshold $_y$ (z-score)						
	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$	Student attendance$_{y+1}$	Functional facilities$_{y+1}$	Math$_{y+1}$	English$_{y+1}$	Urdu$_{y+1}$
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.106	0.048	-0.159	0.242	0.082	0.138
	(0.072)	(0.087)	(0.126)	(0.226)	(0.178)	(0.161)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Mean Dep. Var.	91.6	87.6	90.2	86.9	76.5	84.6
Mean # Times flagged	0.87	1.59	1.61	0.038	0.34	0.065
SD # Times flagged	1.42	2.19	2.81	0.24	0.96	0.33
First stage F-stat	1434.3	1651.3	1721.3	134.4	779.9	395.2
Panel B: Instrument = Dummies for month of close flagging in year y						
	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$	Student attendance$_{y+1}$	Functional facilities$_{y+1}$	Math$_{y+1}$	English$_{y+1}$	Urdu$_{y+1}$
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.085	0.013	-0.245*	0.148	0.229	0.138
	(0.073)	(0.090)	(0.129)	(0.222)	(0.198)	(0.166)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Mean Dep. Var.	91.6	87.6	90.2	86.9	76.5	84.6
Mean # Times flagged	0.87	1.59	1.61	0.038	0.34	0.065
SD # Times flagged	1.42	2.19	2.81	0.24	0.96	0.33
First stage F-stat	177.4	190.0	228.8	21.7	83.3	62.2
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
# Close to the threshold FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table presents the results from estimating equation 3 using alternative instrument definitions. Our main analysis uses as instrument the share of months in a year in which a markaz was flagged close to the threshold. Panel A uses the z-score of the number of times flagged close to the threshold instead of the share. Panel B uses a non-parametric instrument by defining dummies for each month of a year where the markaz was flagged close to the threshold, using May as base dummy. In both panels we control by the number of times the markaz was close to the threshold. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year t on the outcome in year $y+1$. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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 of 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 C3: Intensity of exposure to flagging - non-parametric instrument definition

Panel A: Instrument = Dummy (= 1) for ever flagged close to the threshold in the year						
	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$	Student attendance$_{y+1}$	Functional facilities$_{y+1}$	Math$_{y+1}$	English$_{y+1}$	Urdu$_{y+1}$
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	-0.026	0.310	1.107***	0.603	-1.654*	0.176
	(0.167)	(0.227)	(0.400)	(0.479)	(0.923)	(0.378)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Mean Dep. Var.	91.6	87.6	90.2	86.9	76.5	84.6
Mean # Times flagged	0.87	1.59	1.61	0.038	0.34	0.065
SD # Times flagged	1.42	2.19	2.81	0.24	0.96	0.33
First stage F-stat	215.6	206.2	125.2	20.0	37.4	106.3
Panel B: Instrument = Dummy (= 1) above the median flagging close to the threshold						
	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$	Student attendance$_{y+1}$	Functional facilities$_{y+1}$	Math$_{y+1}$	English$_{y+1}$	Urdu$_{y+1}$
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.033	0.218	0.700*	0.365	0.009	-0.001
	(0.134)	(0.184)	(0.375)	(0.439)	(0.422)	(0.344)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Mean Dep. Var.	91.6	87.6	90.2	86.9	76.5	84.6
Mean # Times flagged	0.87	1.59	1.61	0.038	0.34	0.065
SD # Times flagged	1.42	2.19	2.81	0.24	0.96	0.33
First stage F-stat	213.4	199.0	80.1	13.7	75.0	67.4
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
RDA Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table presents the results from estimating equation 3 using alternative instrument definitions. Panel A uses a dummy for ever flagged close to the threshold in the year. Panel B uses a dummy for above the median of the distribution of flagging close to the threshold. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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 of 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 C4: Impacts of exposure to flagging - non imputed data

Panel A: School outcomes						
	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.236***	0.042	0.384***	0.209	1.223***	-0.336
	(0.050)	(0.116)	(0.055)	(0.166)	(0.089)	(0.316)
N. of obs.	257,592	257,592	257,862	257,862	254,025	254,025
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	90.3	90.3
Mean # Times flagged	0.84	0.84	1.54	1.54	1.56	1.56
SD # Times flagged	1.37	1.37	2.15	2.15	2.76	2.76
First stage F-stat		431.0		405.1		225.9
Panel B: Student scores						
	Dependent variables (range: 0-100)					
	Math$_{y+1}$		English$_{y+1}$		Urdu$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.283***	0.423	1.045***	-0.382	0.284***	-0.208
	(0.080)	(0.348)	(0.138)	(0.375)	(0.083)	(0.339)
N. of obs.	64,643	64,643	64,641	64,641	64,642	64,642
Number markaz	2,418	2,418	2,418	2,418	2,418	2,418
Mean Dep. Var.	86.9	86.9	76.6	76.6	84.6	84.6
Mean # Times flagged	0.037	0.037	0.33	0.33	0.064	0.064
SD # Times flagged	0.23	0.23	0.96	0.96	0.32	0.32
First stage F-stat		43.3		203.0		121.9
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
RDA Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. The aggregated data is based on the non-imputed school-by-month dataset. This table presents the results from estimating equation 3 and is the table equivalent of Panel (a) Figure 3. Panel A reports on schooling outcomes. Panel B reports on student scores. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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 of 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 C5: Intensity of exposure to flagging - balance test maraakiz sample

Panel A: School outcomes						
	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.187***	-0.016	0.043	0.251*	-0.166*	-0.534**
	(0.057)	(0.128)	(0.064)	(0.129)	(0.091)	(0.263)
N. of obs.	204,274	204,274	204,448	204,448	201,650	201,650
Number markaz	2,874	2,874	2,874	2,874	2,874	2,874
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		411.8		512.5		274.1
Panel B: Student scores						
	Dependent variables (range: 0-100)					
	Math$_{y+1}$		English$_{y+1}$		Urdu$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_t$ (z-score)	0.293***	0.511	1.023***	-0.361	0.342***	-0.123
	(0.085)	(0.384)	(0.136)	(0.366)	(0.084)	(0.342)
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		43.1		211.7		122.2
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
RDA Controls	Yes	Yes	Yes	Yes	Yes	Yes

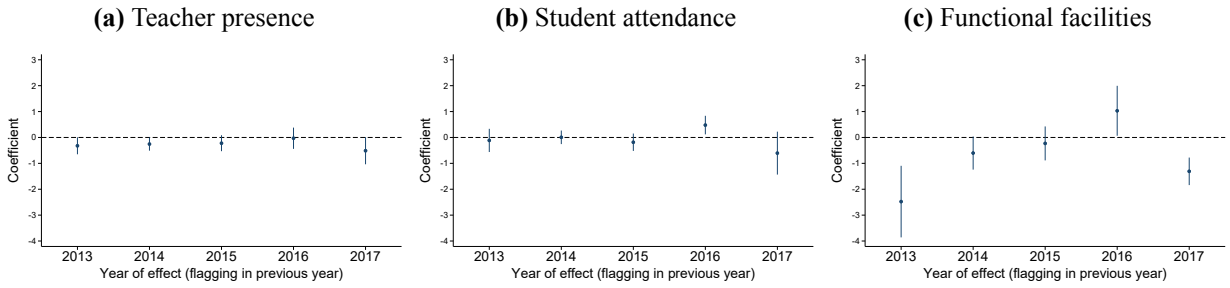
Notes: The unit of analysis is the school-year. This table presents the results from estimating equation 3 on the sample of schools used for the orthogonality analysis of Table C7. Panel A reports on schooling outcomes. Panel B reports on student scores. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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 of 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 C6: Impacts of exposure to flagging - cross-flagging

	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$	Student attendance$_{y+1}$	Functional facilities$_{y+1}$	Math$_{y+1}$	English$_{y+1}$	Urdu$_{y+1}$
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Teacher presence flagging						
# Times flagged$_y$ (z-score)	-0.011	0.104	0.479***	0.008	0.269	0.262
	(0.129)	(0.120)	(0.167)	(0.450)	(0.554)	(0.457)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Panel B: Student attendance flagging						
# Times flagged$_y$ (z-score)	-0.095	0.147	0.158	1.043	0.157	0.443
	(0.140)	(0.134)	(0.192)	(1.044)	(1.437)	(1.180)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Panel C: Functional facilities flagging						
# Times flagged$_y$ (z-score)	-0.349	-0.090	-0.437	0.478	1.954**	1.574**
	(0.220)	(0.212)	(0.292)	(0.661)	(0.811)	(0.768)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Panel D: Math scores flagging						
# Times flagged$_y$ (z-score)	0.554	-0.124	0.093	0.452	1.046**	0.531
	(0.344)	(0.251)	(0.206)	(0.394)	(0.454)	(0.402)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Panel E: English scores flagging						
# Times flagged$_y$ (z-score)	0.311	0.021	0.303*	-0.018	-0.231	0.152
	(0.228)	(0.203)	(0.173)	(0.302)	(0.366)	(0.313)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
Panel F: Urdu scores flagging						
# Times flagged$_y$ (z-score)	-0.115	-0.140	0.100	-0.018	-0.181	-0.125
	(0.191)	(0.179)	(0.192)	(0.318)	(0.341)	(0.342)
N. of obs.	257,592	257,865	254,058	67,385	67,383	67,384
Number markaz	3,567	3,567	3,567	2,721	2,721	2,721
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table presents the results from estimating equation 3 on outcomes different to the one causing the flagging. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at title of the panel. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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 of the share of correct answers in standardized exams performed on a random sample of primary students. Standard errors clustered by markaz in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Figure C1: Impacts of exposure to flagging - effect by year



Note: This figure presents the results from estimating equation 3 by separating the endogenous variable and instrument by year, and using as base period 2012, the first year of the command-and-control system implementation. The figure test robustness for the results reported in Table C1 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 standard deviation increase in flagging.

C.2 Identifying Assumptions

Our instrument consist on the aggregation of monthly RD shocks. First, we test for the (quasi) random assignment assumption by examining if the treatment variable and instrument correlate with past outcomes. We estimate equation 1 on lagged ($t-1$) outcomes using as explanatory variables the number of times flagged (D) and number of times flagged while being in the threshold sample (Z). We use the number of times flagged close to the threshold instead of the share of times (as in our main analysis) to have comparable units between the instrument and the endogenous variable.

Table C7 shows the results from equation 1 including the RDA controls. Panel A reports on 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. Columns (1)-(3) show when using the number of times flagged (D) as an explanatory variable that all coefficients are significant but smaller than one percentage point. The results might reflect that maraakiz performing unusually well in $t-1$ may be more likely to be flagged in t if performance falls back the following year. It can also signal that unusually high performance is difficult to sustain, so that maraakiz with high outcomes in $t-1$ are more likely to be flagged in t. In any scenario, the significance of the coefficients indicates that the treatment is likely endogenous and correlated with previous performance or other school characteristics predicting flagging. In contrast, columns (4)-(6) for the number of times flagged close to the threshold (Z) show an order of magnitude smaller and non-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 performance from those randomly not exposed to it.[k]

Second, we test for manipulation of flagging assignment in Figure C2. We observe a smooth distribution around the threshold for all 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. For the remaining five outcomes we do not observe significant jumps at the threshold, and for all outcomes the discontinuity test is not significant. Overall, we find no evidence for manipulation.

Table C8 test for the relevance assumption by displaying the first stage coefficients from equation 2. We normalize the instrument in standard deviations to ease interpretation. The table show that a one standard deviation increase in exposure to flagging close to the threshold induces an increase

[k] 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 C5 and discuss it below.

Table C7: Orthogonality of the instrument

Panel A: Markaz average schooling outcomes

	Dependent variables (range: 0-100)					
	Teacher presence$_{y-1}$	Student attendance$_{y-1}$	Functional facilities$_{y-1}$	Teacher presence$_{y-1}$	Student attendance$_{y-1}$	Functional facilities$_{y-1}$
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ – D (z-score)	0.344**	0.264***	-0.840***			
	(0.146)	(0.087)	(0.168)			
# Times flagged threshold$_y$ – Z (z-score)				0.099	-0.047	0.196
				(0.071)	(0.082)	(0.160)
N. of obs.	7,478	7,478	7,478	7,478	7,478	7,478
Number markaz	2,874	2,874	2,874	2,874	2,874	2,874
Mean Dep. Var	92.0	88.8	93.7	92.0	88.8	93.7
Mean # Times flagged	0.21	0.34	0.22	0.21	0.34	0.22
SD # Times flagged	0.56	0.76	0.71	0.56	0.76	0.71

Panel B: Markaz average student scores

	Dependent variables (range: 0-100)					
	Math$_{y-1}$	English$_{y-1}$	Urdu$_{y-1}$	Math$_{y-1}$	English$_{y-1}$	Urdu$_{y-1}$
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ – D (z-score)	0.498***	0.712***	0.405***			
	(0.144)	(0.106)	(0.082)			
# Times flagged threshold$_y$ – Z (z-score)				0.087	-0.022	-0.083
				(0.084)	(0.125)	(0.086)
N. of obs.	3,674	3,674	3,674	3,674	3,674	3,674
Number markaz	1,837	1,837	1,837	1,837	1,837	1,837
Mean Dep. Var	86.9	75.9	84.2	86.9	75.9	84.2
Mean #Times flagged	0.038	0.28	0.050	0.038	0.28	0.050
SD # Times flagged	0.22	0.61	0.24	0.22	0.61	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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the markaz-year. This table shows the results from estimating equation 1 including RDA controls. Outcomes in the top of each column, measured in year $y-1$ in scale from 0 to 100. Panel A reports on schooling outcomes. Panel B reports on student scores. # Times flagged$_y$ counts the number of times a markaz was flagged in year y in the outcome reported at the top of the column. # Times flagged threshold$_y$ counts the number of times flagged while being close to the flagging threshold in the outcome reported at the top. We use the number of times flagged close to the threshold instead of the share of times (as in our main analysis) to have comparable units between the instrument and the endogenous variable. The variables for the # Times flagged are normalized (z-score). Following Borusyak and Kolerman-Shemer (2024), the RDA controls are defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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 as 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year t. Standard errors clustered by markaz are in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

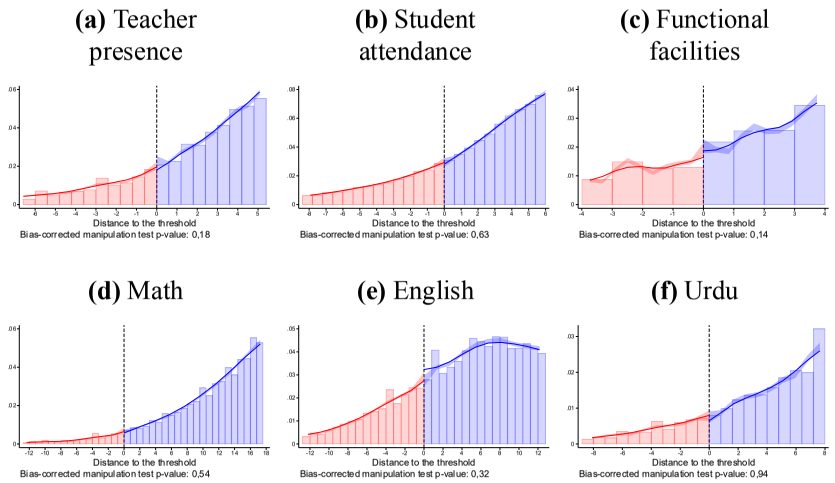
in overall flagging ranging from 0.404 to 0.565 standard deviations, which account for one additional flag. Finally, Figure C3 test for the monotonicity assumption and shows a positive and monotonic relationship between the instrument and the endogenous variable, residualized from the fixed effects reported in equation 2, suggesting that the monotonicity assumption is met.

Table C8: First stage

	Dependent variable: # Times flagged$_y$ (z-score) in the variable below					
	Teacher presence	Student attendance	Functional facilities	Math	English	Urdu
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged close to threshold $_y$ (z-score)	0.465***	0.446***	0.373***	0.322***	0.485***	0.379***
	(0.021)	(0.019)	(0.022)	(0.050)	(0.033)	(0.035)
N. of obs.	258,147	258,147	258,147	258,147	258,147	258,147
Number markaz	3,567	3,567	3,567	3,567	3,567	3,567
Mean Endog. Var.	0.0021	0.098	0.094	-0.13	-0.26	-0.16
SD Endog. Var.	1.01	1.05	1.07	0.62	0.71	0.65
Mean Instrument	0.073	0.15	0.082	-0.068	-0.21	-0.10
SD Intrument	1.07	1.12	1.13	0.66	0.73	0.68
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
RDA Controls	Yes	Yes	Yes	Yes	Yes	Yes

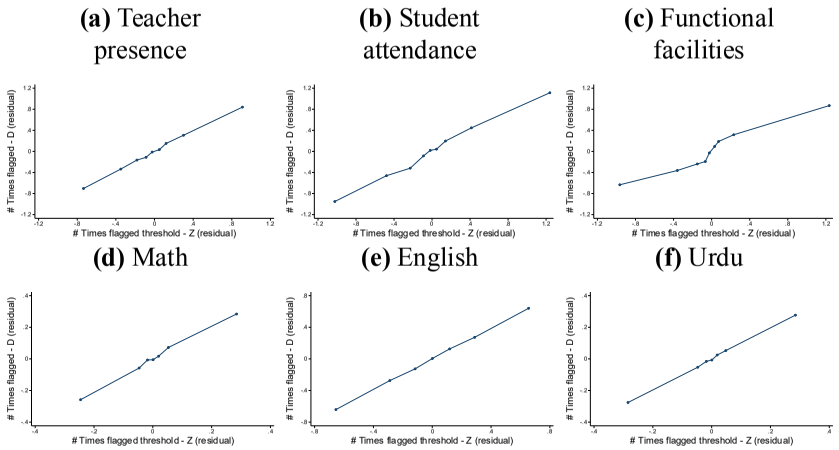
Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 2. The outcome # Times flagged $_y$ counts the number of times an schools is in a flagged markaz on the performance variable reported in the top of each column. # Times flagged close to threshold$_y$ counts the number of times a markaz was flagged while being close to the threshold in the outcome reported at the top of the column. The variables are normalized (z-score). Following Borusyak and Kolerman-Shemer (2024), the RDA controls are defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. Year is measured as school-year (September to May). *Mean Endog. Var* and *SD Endog. Var* indicate the mean and standard deviation of # Times flagged$_y$. *Mean Instrument* and *SD Instrument* indicate the mean and standard deviation of # Times flagged close to threshold$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. Standard errors clustered by markaz are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure C2: 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 C3: 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.

C.3 Alternative Margins of School Response

The main results show negligible effects of flagging on yearly averages of school performance relying on the red-flagging threshold to measure flagging exposure. In this sub-section we explore alternative outcomes that measure other performance margins and heterogeneity by school characteristics that might reflect how flagging is being translated into changes in school performance. We also explore other flagging thresholds and hierarchical levels around which incentives for high-performance might be stronger.

Alternative performance measures: best and worst annual school performance We showed that there is no evidence that more intense flagging improved outcomes on average. However, it could be the case that the intensity of command-and-control accountability worked in improving the worst or best months of performance in a given school year. We find no evidence for this in Table C9, 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, all but two (close to one percentage point and non- or barely significant) RDA estimate show that improvements remain smaller than a percentage point, suggesting that a more intense exposure to accountability in the command-and-control scheme did not substantively improve outcomes across the school year.

Targeting and heterogeneous responses by best and worst performing schools We examine if, despite the average null, specific schools improved their performance by exploring heterogeneity by school characteristics, considering that administrative responses to a more intense exposure to flagging could result in AEOs targeting schools differentially. We use the fact that whether a markaz is flagged and subject to reprimands and bureaucratic action 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 C10 reports minimal effects in both types of schools, suggesting that a more intense accountability did not move outcomes differentially in best or worst schools in maraakiz.

Change of incentives around higher performance: green/orange flagging 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 under-performance of the schools in their markaz. Yet, it might be that stronger accountability responses were generated around higher flagging thresholds, where bureaucrats can be rewarded due to good

performance. We test in Table C11 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: district rankings 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 y. We find small effects in Table C12, suggesting that additional accountability pressures also did not produce any movement in school outcomes. 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.

Effects may also 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. Table C13 reports the results. Due to the small sample size is harder to detect an effect, but most coefficients remain below one percentage point or are negative for bottom and top-ranked districts.

Political pressure on improved performance: political alignment with ruling party 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. Because maraakiz might overlap across constituencies, we also define an indicator for not fully aligned markaz. Table C14 shows the results of the interaction of a school alignment indicator with the number of times flagged. The RDA estimates reveal no statistically discernible heterogeneity by this measure of political alignment. This result does not rule out other political channels, but provides no evidence that

the effectiveness of flag-triggered accountability differed by constituency alignment with the Chief Minister's party.

Table C9: Intensity of exposure to flagging - other performance metrics

Panel A: Best performance in the year as outcome

	Dependent variables (range: 0-100)											
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$		Math$_{y+1}$		English$_{y+1}$		Urdu$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.090***	-0.051	0.30***	-0.024	1.02***	-0.25	-0.16*	0.73	0.34**	-0.27	-0.059	0.53
	(0.034)	(0.080)	(0.048)	(0.11)	(0.078)	(0.21)	(0.091)	(0.54)	(0.16)	(0.44)	(0.093)	(0.38)
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		494.2		577.0		288.5		39.5		220.3		122.4

Panel B: Worst performance in the year as outcome

	Dependent variables (range: 0-100)											
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$		Math$_{y+1}$		English$_{y+1}$		Urdu$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_t$ (z-score)	0.372***	0.097	0.690***	0.472	0.958***	-1.507	1.134***	-0.453	1.695***	0.036	0.901***	-1.025*
	(0.123)	(0.337)	(0.142)	(0.330)	(0.211)	(1.318)	(0.134)	(0.583)	(0.192)	(0.531)	(0.138)	(0.524)
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		494.2		577.0		288.5		39.5		220.3		122.4
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
RDA Controls	Yes	Yes	Yes	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 on alternative performance measures, different from the average annual performance as used for the main results of Panel (a) Figure 3. Panel A report for each outcome best monthly performance in a school year. Panel B reports for the worst monthly performance. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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 of 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 C10: Intensity of exposure to flagging by school performance

Panel A: Effects on best performance schools

	Dependent variables (range: 0-100)											
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$		Math$_{y+1}$		English$_{y+1}$		Urdu$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	-0.047***	-0.000	-0.145***	0.067	1.042***	-0.427	0.014	-0.408**	0.201**	-0.297	0.025	-0.141
	(0.015)	(0.043)	(0.021)	(0.045)	(0.120)	(0.395)	(0.045)	(0.174)	(0.089)	(0.228)	(0.045)	(0.138)
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		487.2		554.2		252.2		86.8		278.4		173.8

Panel B: Effects on worst performing schools

	Dependent variables (range: 0-100)											
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$		Math$_{y+1}$		English$_{y+1}$		Urdu$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_t$ (z-score)	0.383***	-0.004	0.745***	0.069	1.351***	-0.247	0.205**	0.505	0.605***	-0.066	0.312***	0.206
	(0.076)	(0.163)	(0.062)	(0.149)	(0.117)	(0.384)	(0.088)	(0.505)	(0.137)	(0.378)	(0.088)	(0.349)
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		414.5		475.2		245.0		28.5		164.2		97.6
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 on samples of schools separated by median performance within the markaz to test if conclusions from Panel (a) Figure 3 change under specific school characteristics. Outcomes in the top of each column, measured in year $y+1$ 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. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by the its share of months. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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 of 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 C11: Impacts of exposure to flagging - orange flagging

Panel A: School outcomes						
	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.150***	-0.061	0.058	0.167*	0.211***	-0.207
	(0.044)	(0.103)	(0.053)	(0.086)	(0.054)	(0.174)
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		603.7		1521.9		171.8
Panel B: Student scores						
	Dependent variables (range: 0-100)					
	Math$_{y+1}$		English$_{y+1}$		Urdu$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.408***	0.122	0.181	0.544	0.499***	-0.019
	(0.091)	(0.293)	(0.147)	(0.526)	(0.115)	(0.317)
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		195.3		124.8		258.2
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 defining flagging variables 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. Panel A report the results on school outcomes. Panel B report the results on student test-scores. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by the its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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 as 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 C12: Intensity of exposure to flagging - heterogeneity by district ranking

	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	-0.057	-0.321	0.907***	0.681**	1.351***	-0.734
	(0.123)	(0.272)	(0.129)	(0.320)	(0.219)	(0.458)
Bottom district$_y$ × # Times flagged$_y$ (z-score)	0.327***	0.362	-0.496***	-0.664**	0.055	0.542
	(0.126)	(0.267)	(0.136)	(0.312)	(0.237)	(0.449)
Top district$_y$ × # Times flagged$_y$ (z-score)	0.203	0.303	-0.572***	-0.474	-0.305	0.083
	(0.128)	(0.268)	(0.136)	(0.320)	(0.234)	(0.424)
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 adding an additional interaction for the district ranking to test for heterogeneity of the main results of Panel (a) in Figure 3. For the RDA estimations using 3, we first instrument the number of times flagged using equation 2, then interact the predicted variable with the district ranking indicators. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. Bottom district equals one if the average position of the district is in the bottom. Top district equals one if the average position is in the top five, relative to the districts outside the bottom and top positions. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by the its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table C13: Impacts of exposure to flagging - district ranking

	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
District in the bottom 5	-0.207		0.366		2.209	
	(0.456)		(0.628)		(1.367)	
District in the top 5		0.614		-0.156		0.780
		(0.428)		(0.702)		(0.873)
N. of obs.	360	360	360	360	360	360
Number districts	36	36	36	36	36	36
Mean Dep. Var.	90.7	90.7	87.8	87.8	91.8	91.8
First stage F-stat	279.0	670.3	279.0	670.3	279.0	670.3
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)$_y$ is a dummy equal to one if the district in a quarter y was in the bottom of top of the district ranking, instrumented by dummy variables equal to one if the district in a quarter y 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^{nd} or 33^{rd} position. We control by District in the ranking threshold$_y$, 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^{rd} position for the bottom. We include district and quarter fixed effects. Outcomes in the top of each column, measured in quarter $y+1$ 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 y. 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 C14: Intensity of exposure to flagging - heterogeneity by political alignment

	Dependent variables (range: 0-100)					
	Teacher presence$_{y+1}$		Student attendance$_{y+1}$		Functional facilities$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)
# Times flagged$_y$ (z-score)	0.071	0.166	0.527***	-0.105	1.209***	-0.436
	(0.120)	(0.194)	(0.129)	(0.259)	(0.226)	(0.456)
Not fully aligned × # Times flagged$_y$ (z-score)	0.057	-0.245	-0.244*	0.126	-0.010	0.113
	(0.123)	(0.184)	(0.143)	(0.259)	(0.255)	(0.418)
Fully aligned × # Times flagged$_y$ (z-score)	0.212*	-0.166	-0.052	0.422	0.048	-0.114
	(0.121)	(0.184)	(0.137)	(0.258)	(0.251)	(0.407)
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 adding an additional interaction for the political alignment to test for heterogeneity of the main results of Panel (a) in Figure 3. 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. For the RDA estimations using 3, we first instrument the number of times flagged using equation 2, then interact the predicted variable with the alignment indicators. Outcomes in the top of each column, measured in year $y+1$ in scale from 0 to 100. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. 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. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by the its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. 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.4 Changes in the Wider 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. In this sub-section, we use data on the administrative machinery of government to assess whether there was any change in bureaucratic activity due to the oversight intervention.

C.4.1 Bureaucratic Oversight Visits, Transfers and Postings

A natural response by public officials repeatedly flagged would be to visit poorly performing schools to undertake remedial work. School visits are a standard part of the AEO work and a mechanism to resolve issues that schools face in functioning effectively. Public officials can also intervene in school management by moving head teachers in response to flagging. We explore whether a higher intensity of accountability affected these two margins estimating equation 3 using as outcomes (i) the share of months that a school received a bureaucratic visit in a year and (ii) the share of months in which the school reported a different head-teachers.

Panel A of Table C15 indicates that more intense flagging has small (all but one non-significant and lower than a percentage point) effects on bureaucratic visits. Panel B shows no significant or small effects on head teacher transfers and postings. The results are robust when using the linear (number of months with close-flagging) and non-parametric (dummies for the months with close-flagging) definitions of the instrument (Table C16). The results are also robust to not imputing the missing data for non-observed school-months before aggregating the monthly data-packs at the school-by-year level (Table C17).

In addition, we test for bureaucratic incentives at higher levels of the hierarchy measuring district manager rotation. We aggregate our data at the district-quarter level to run a regression of the manager being in the top/bottom parts of the district ranking in a quarter on whether there was rotation of district officer in the following quarter. Table C18 does not show statistically significant effects. The coefficient on the bottom districts is large relative to the baseline rotation probability, but it is imprecisely estimated. The p-values from a wild-cluster bootstrap test of the null that the coefficient equals zero are likewise not statistically significant. Given the small district-quarter sample, we do not detect a clear pattern of transfers, but rotation incentives for district officers in the worst-ranked districts cannot be ruled out.

C.4.2 Budget Allocation and Utilization

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 it comes from the government. The rest comes from non-government sources, consisting on local community bodies that are authorized to raise funds from parents, philanthropists, and other non-government sources. District officials have limited influence over non-government funding, but 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 C19 shows the results on the total development funds and its expenditures. The RDA results show 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 roughly 6 to 21 percent. However, the increase is minor: for the *average* school this amounts to a maximum of around 70 additional dollars annually (9.5 thousand Pakistan rupees, using conversion rate from December 2018 – 139 rupees per dollar). The effect is not statistically significant, but the increase arises solely from government funds (see Table C20). We do not observe an effect on resource availability arising from more intensive accountability for lapses in functional facilities and student scores variables.

For expenditures, teacher presence flagging coefficient suggests that a one standard deviation increase in flagging increases expenditure by 18 percent (s.e.=10.7) significant only at the 10%. The coefficient for flagging on student attendance suggests barely significant increase of 18 percent (s.e.=14). Yet, both coefficients are translated to small monetary values, thus economically negligible. Thus, our microdata does not provide evidence that the increase in funding and expenditures translated into significant improvement in school functioning, which is in line with the idea that beyond senior management action, a more intense command-and-control accountability has limited effects on bureaucratic activity.

Table C21 estimate the effects on budget variables using alternative instrument definitions. The coefficient for teacher presence decrease increase to 16% and for student attendance decrease to 8-11%. In both cases the monetary value is still small. Table C22 show the results not imputing the missing data before aggregating the monthly data-packs at the school-by-year level. The results

are similar and hold across versions.

Table C15: 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$_{y+1}$											
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	1.117***	-0.663	1.779***	0.880*	0.072	0.112	0.134	2.012*	0.127	-0.757	0.302	-0.140
	(0.206)	(0.459)	(0.249)	(0.478)	(0.233)	(0.750)	(0.210)	(1.037)	(0.348)	(0.860)	(0.217)	(0.842)
N. of obs.	257,231	257,231	257,504	257,504	253,699	253,699	67,263	67,263	67,261	67,261	67,262	67,262
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.	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9
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		493.8		577.1		288.3		39.4		220.2		122.9
Panel B: Transfers and postings												
Flagging variable	Teacher presence		Student attendance		Functional facilities		Math		English		Urdu	
	Dependent variable (range: 0-100) Change head teacher$_{y+1}$											
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.101	-0.381	-0.006	-0.437	0.429***	-0.896*	-0.116	0.214	-0.108	0.562	-0.005	-0.122
	(0.100)	(0.248)	(0.117)	(0.281)	(0.112)	(0.464)	(0.100)	(0.470)	(0.135)	(0.386)	(0.095)	(0.323)
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		494.2		577.0		288.5		39.5		220.3		122.4
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 on bureaucratic effort outcomes. Panel A report on the bureaucratic oversight dependent variable measured as the share of oversight visit done by the AEO in year $y+1$. Panel B reports on the transfers and posting dependent variable, measured as the share of schools that had a change of head teacher in year $y+1$. Both panels measure the outcomes in scale 0-100. Variables in the top of each column indicate the flagging variable. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. Standard errors clustered by markaz in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table C16: Intensity of exposure to flagging - effect on AEO's effort - alternative instrument

Panel A: Bureaucratic oversight												
Flagging variable	Teacher presence		Student attendance		Functional facilities		Math		English		Urdu	
	Dependent variable (range: 0-100) Visited schools$_{y+1}$											
Instrument variation	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	-0.127	0.105	0.626**	0.578*	-0.135	-0.197	1.484*	1.087	-0.390	-1.014	-0.351	-0.409
	(0.272)	(0.278)	(0.291)	(0.305)	(0.330)	(0.322)	(0.828)	(0.756)	(0.602)	(0.682)	(0.566)	(0.541)
N. of obs.	257,786	257,786	257,786	257,786	257,786	257,786	257,786	257,786	257,786	257,786	257,786	257,786
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.	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9	91.9
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	1436.7	177.6	1651.6	190.1	1734.1	229.7	137.6	22.1	786.9	82.4	418.9	65.9
Panel B: Transfers and postings												
Flagging variable	Teacher presence		Student attendance		Functional facilities		Math		English		Urdu	
	Dependent variable (range: 0-100) Change head teacher$_{y+1}$											
Instrument variation	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.239*	0.396***	-0.158	-0.144	0.101	0.112	0.692*	0.395	0.319	0.114	-0.491**	-0.537**
	(0.145)	(0.152)	(0.162)	(0.170)	(0.187)	(0.181)	(0.384)	(0.379)	(0.275)	(0.307)	(0.233)	(0.226)
N. of obs.	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147
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.	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	1436.6	177.5	1651.9	190.1	1734.4	229.5	137.8	22.1	787.9	82.5	418.8	65.9
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
# Close to the threshold FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 on bureaucratic effort outcomes using alternative instrument definitions. Our main analysis uses as instrument the share of months in a year in which a markaz was flagged close to the threshold. *Linear* columns use the z-score of the number of times flagged close to the threshold instead of the share. *Non-param* columns uses a non-parametric instrument by defining dummies for each month of a year where the markaz was flagged close to the threshold, using May as base dummy. We control by the number of times the markaz was close to the threshold instead of the RDA controls. Panel A report on the bureaucratic oversight dependent variable measured as the share of oversight visit done by the AEO in year $y+1$. Panel B reports on the transfers and posting dependent variable, measured as the share of schools that had a change of head teacher in year $y+1$. Both panels measure the outcomes in scale 0-100. Variables in the top of each column indicate the flagging variable. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. Standard errors clustered by markaz in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table C17: Intensity of exposure to flagging - effect on AEO's effort non-imputed data

Panel A: Bureaucratic oversight												
Flagging variable	Teacher presence		Student attendance		Functional facilities		Math		English		Urdu	
	Dependent variable (range: 0-100) Visited schools$_{y+1}$											
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	1.031***	-0.272	1.399***	-0.040	0.009	0.672	0.122	1.460	0.291	-0.909	0.267	-0.675
	(0.190)	(0.445)	(0.231)	(0.542)	(0.219)	(0.827)	(0.217)	(1.061)	(0.388)	(0.936)	(0.229)	(0.846)
N. of obs.	254,225	254,225	254,349	254,349	250,641	250,641	63,868	63,868	63,867	63,867	63,867	63,867
Number markaz	3,561	3,561	3,561	3,561	3,561	3,561	2,416	2,416	2,416	2,416	2,416	2,416
Mean Dep. Var.	92.3	92.3	92.3	92.3	92.3	92.3	92.3	92.3	92.3	92.3	92.3	92.3
Mean # Times flagged	0.84	0.84	1.54	1.54	1.56	1.56	0.037	0.037	0.33	0.33	0.064	0.064
SD # Times flagged	1.37	1.37	2.15	2.15	2.76	2.76	0.23	0.23	0.96	0.96	0.32	0.32
First stage F-stat		431.7		401.6		226.7		46.3		201.8		122.1
Panel B: Transfers and postings												
Flagging variable	Teacher presence		Student attendance		Functional facilities		Math		English		Urdu	
	Dependent variable (range: 0-100) Change head teacher$_{y+1}$											
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.059	-0.084	-0.044	-0.150	0.330***	-0.823**	-0.040	-0.027	-0.070	0.097	0.008	-0.214
	(0.074)	(0.190)	(0.087)	(0.262)	(0.086)	(0.374)	(0.064)	(0.292)	(0.095)	(0.265)	(0.062)	(0.215)
N. of obs.	257,592	257,592	257,862	257,862	254,025	254,025	64,643	64,643	64,641	64,641	64,642	64,642
Number markaz	3,567	3,567	3,567	3,567	3,567	3,567	2,418	2,418	2,418	2,418	2,418	2,418
Mean Dep. Var.	8.32	8.32	8.32	8.32	8.32	8.32	8.32	8.32	8.32	8.32	8.32	8.32
Mean # Times flagged	0.84	0.84	1.54	1.54	1.56	1.56	0.037	0.037	0.33	0.33	0.064	0.064
SD # Times flagged	1.37	1.37	2.15	2.15	2.76	2.76	0.23	0.23	0.96	0.96	0.32	0.32
First stage F-stat		431.0		405.1		225.9		43.3		203.0		121.9
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. The aggregated data is based on the non-imputed school-by-month dataset. This table shows the results from estimating equation 3 on bureaucratic effort outcomes. Panel A report on the bureaucratic oversight dependent variable measured as the share of oversight visit done by the AEO in year $y+1$. Panel B reports on the transfers and posting dependent variable, measured as the share of schools that had a change of head teacher in year $y+1$. Both panels measure the outcomes in scale 0-100. Variables in the top of each column indicate the flagging variable. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year y. Standard errors clustered by markaz in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table C18: Impacts of exposure to flagging - district officer rotation

	Pr(rotation district officer)	
	(1)	(2)
District in the bottom 5	0.092	
	(0.163)	
	[0.163]	
District in the top 5		-0.027
		(0.137)
		[0.124]
N. of obs.	360	360
Number districts	36	36
Mean Dep. Var.	0.23	0.23
First stage F-stat	279.0	670.3
Wild bootstrap p-value	0.605	0.858
District FE	Yes	Yes
Time FE	Yes	Yes
District in the ranking threshold$_t$ FE	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. Results from estimating a modified version of equation 3: District in the top (bottom)$_y$ is a dummy equal to one if the district in a quarter y was in the bottom of top of the district ranking, instrumented by dummy variables equal to one if the district in a quarter y 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^{nd} or 33^{rd} position. We control by District in the ranking threshold$_y$, 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^{rd} position for the bottom. We include district and quarter fixed effects. Outcomes in the top equals one if in quarter $y+1$ there was a change of district officer. *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 y. Standard errors clustered by district in parentheses. Bootstrapped standard errors from 10 thousand replications in brackets. Wild bootstrap reports p-values from a district-level wild-cluster bootstrap test of the null that the coefficient equals zero, using Webb weights and 9,999 replications. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table C19: 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$_{y+1}$						Expenditure$_{y+1}$					
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.170***	0.063	0.187***	0.216	-0.016	-0.046	0.066	0.183*	0.111**	0.183	0.056	-0.243
	(0.048)	(0.141)	(0.057)	(0.184)	(0.048)	(0.282)	(0.040)	(0.107)	(0.045)	(0.140)	(0.045)	(0.219)
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		494.2		577.0		288.5		494.2		577.0		288.5
Panel B: Student scores flagging												
Flagging variable	Math		English		Urdu		Math		English		Urdu	
	Dependent variables (logs)											
	Total Funds$_{y+1}$						Expenditure$_{y+1}$					
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.052*	-0.055	-0.036	0.020	0.039	0.054	0.059**	-0.027	0.030	0.070	0.044**	0.002
	(0.027)	(0.107)	(0.034)	(0.099)	(0.024)	(0.097)	(0.027)	(0.100)	(0.031)	(0.085)	(0.020)	(0.068)
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		39.5		220.3		122.4		39.5		220.3		122.4
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 on the development funds received by the school and expenditure, both are measured in log scale in year $y+1$. Panel A report the effect on budget variables from schooling outcomes flagging. Panel B reports on the student scores flagging. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *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 C20: 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$_{y+1}$		Non Government Funds$_{y+1}$		Government Funds$_{y+1}$		Non Government Funds$_{y+1}$		Government Funds$_{y+1}$		Non Government Funds$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.163***	0.039	0.081***	-0.097	0.228***	0.178	0.005	0.212**	-0.056	-0.023	-0.023	0.057
	(0.051)	(0.153)	(0.031)	(0.079)	(0.060)	(0.191)	(0.036)	(0.092)	(0.050)	(0.293)	(0.037)	(0.155)
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		494.2		494.2		577.0		577.0		288.5		288.5

Panel B: Student scores flagging

Flagging variable	Math				English				Urdu			
	Dependent variables (range: 0-100)											
	Government Funds$_{y+1}$		Non Government Funds$_{y+1}$		Government Funds$_{y+1}$		Non Government Funds$_{y+1}$		Government Funds$_{y+1}$		Non Government Funds$_{y+1}$	
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.037	0.015	0.043*	-0.014	-0.040	-0.009	-0.026	-0.066	0.029	0.071	0.058**	0.148
	(0.025)	(0.099)	(0.022)	(0.107)	(0.032)	(0.092)	(0.034)	(0.094)	(0.022)	(0.088)	(0.024)	(0.093)
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		39.5		39.5		220.3		220.3		122.4		122.4
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 on the the disaggregation of development funds to test the source of the effects observed in Table C19. Panel A report the effect on budget variables from schooling outcomes flagging. Panel B reports on the student scores flagging. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *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 C21: Intensity of exposure to flagging - effect on budget – alternative instruments

Panel A: School outcomes flagging												
Flagging variable	Teacher presence		Student attendance		Functional facilities		Teacher presence		Student attendance		Functional facilities	
	Dependent variables (logs)											
	Total Funds$_{y+1}$						Expenditure$_{y+1}$					
Instrument variation	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.168**	0.164**	0.087	0.119	-0.056	-0.039	0.046	0.065	-0.030	-0.032	-0.054	0.006
	(0.085)	(0.082)	(0.104)	(0.108)	(0.102)	(0.102)	(0.068)	(0.070)	(0.075)	(0.080)	(0.082)	(0.081)
N. of obs.	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147	258,147
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												
Panel B: Student scores flagging												
Flagging variable	Math		English		Urdu		Math		English		Urdu	
	Dependent variables (logs)											
	Total Funds$_{y+1}$						Expenditure$_{y+1}$					
Instrument variation	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.	Linear	Non-param.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.019	0.025	-0.010	-0.005	-0.027	-0.008	0.050	0.099	-0.020	-0.047	-0.016	-0.012
	(0.063)	(0.065)	(0.051)	(0.053)	(0.047)	(0.046)	(0.061)	(0.064)	(0.042)	(0.048)	(0.033)	(0.032)
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												
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
# Close to the threshold FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. This table shows the results from estimating equation 3 on the development funds received by the school and expenditure, both measure in log scale in year $y+1$ using alternative instrument definitions. Our main analysis uses as instrument the share of months in a year in which a markaz was flagged close to the threshold. *Linear* columns use the z-score of the number of times flagged close to the threshold instead of the share. *Non-param* columns uses a non-parametric instrument by defining dummies for each month of a year where the markaz was flagged close to the threshold, using May as base dummy. We control by the number of times the markaz was close to the threshold instead of the RDA controls. Panel A report the effect on budget variables from schooling outcomes flagging. Panel B reports on the student scores flagging. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *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 C22: Intensity of exposure to flagging - effect on budget non-imputed data

Panel A: School outcomes flagging												
Flagging variable	Teacher presence		Student attendance		Functional facilities		Teacher presence		Student attendance		Functional facilities	
	Dependent variables (logs)											
	Total Funds$_{y+1}$						Expenditure$_{y+1}$					
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.156***	0.055	0.158***	0.170	-0.023	-0.043	0.058	0.188*	0.077*	0.154	0.067	-0.251
	(0.047)	(0.143)	(0.054)	(0.217)	(0.047)	(0.315)	(0.039)	(0.104)	(0.044)	(0.152)	(0.045)	(0.250)
N. of obs.	257,592	257,592	257,862	257,862	254,025	254,025	257,592	257,592	257,862	257,862	254,025	254,025
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	0.84	0.84	1.54	1.54	1.56	1.56	0.84	0.84	1.54	1.54	1.56	1.56
SD # Times flagged	1.37	1.37	2.15	2.15	2.76	2.76	1.37	1.37	2.15	2.15	2.76	2.76
First stage F-stat		431.0		405.1		225.9		431.0		405.1		225.9
Panel B: Student scores flagging												
Flagging variable	Math		English		Urdu		Math		English		Urdu	
	Dependent variables (logs)											
	Total Funds$_{y+1}$						Expenditure$_{y+1}$					
	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA	OLS	RDA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
# Times flagged$_y$ (z-score)	0.050**	-0.026	-0.031	0.031	0.040*	0.053	0.052**	0.002	0.029	0.080	0.042**	0.002
	(0.025)	(0.092)	(0.034)	(0.100)	(0.023)	(0.092)	(0.025)	(0.085)	(0.030)	(0.085)	(0.019)	(0.064)
N. of obs.	64,643	64,643	64,641	64,641	64,642	64,642	64,643	64,643	64,641	64,641	64,642	64,642
Number markaz	2,418	2,418	2,418	2,418	2,418	2,418	2,418	2,418	2,418	2,418	2,418	2,418
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	0.037	0.037	0.33	0.33	0.064	0.064	0.037	0.037	0.33	0.33	0.064	0.064
SD # Times flagged	0.23	0.23	0.96	0.96	0.32	0.32	0.23	0.23	0.96	0.96	0.32	0.32
First stage F-stat		43.3		203.0		121.9		43.3		203.0		121.9
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
RDA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The unit of analysis is the school-year. The aggregated data is based on the non-imputed school-by-month dataset. This table shows the results from estimating equation 3 on the overall amount of development funds received by the school and the overall expenditure, both measure in log scale in year $y+1$. Panel A report the effect on budget variables from schooling outcomes flagging. Panel B reports on the student scores flagging. # Times flagged$_y$ (z-score) is the standardized number of times a markaz was flagged in year y in the outcome reported at the top of the column. The coefficients are interpreted as the effect of a one standard deviation in # Times flagged in year y on the outcome in year $y+1$. OLS columns show the results from equation 1. RDA columns show the results from equation 3 after instrumenting the # Times flagged by the share of times flagged close to the threshold in a school year. Following Borusyak and Kolerman-Shemer (2024), all columns also include RDA controls defined as (i) the share of months in the school year in which the markaz was close to the threshold, (ii) the average running variable in months close to the threshold, weighted by the share of months close to the threshold, and (iii) the average running variable in close to the threshold flagging months, weighted by its share of months. All columns also include markaz and year fixed effects, and district linear time trends. 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$_y$. *Mean. Dep. Var* shows the average outcome in the markaz in year t. Standard errors clustered by markaz in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

D Naive Evaluations of Response

We study if the limited impact of flagging is driven by a reversion to the mean in Section 4. We use the school-by-month data to implement a stacked difference-in-discontinuities analysis comparing monthly outcomes for schools in flagged and non-flagged maraakiz, before and after the month when a flag occurs. Our design allow us to estimate the effects for the full and 'threshold sample'. The latter is defined as schools in markaz barely above and below the threshold. In this section we provide further details of the results as well as robustness tests. We restrict the estimation to units within the optimal bandwidth using the same procedure as in Appendix C, allowing for a quadratic polynomial.[1]

D.1 Additional Results

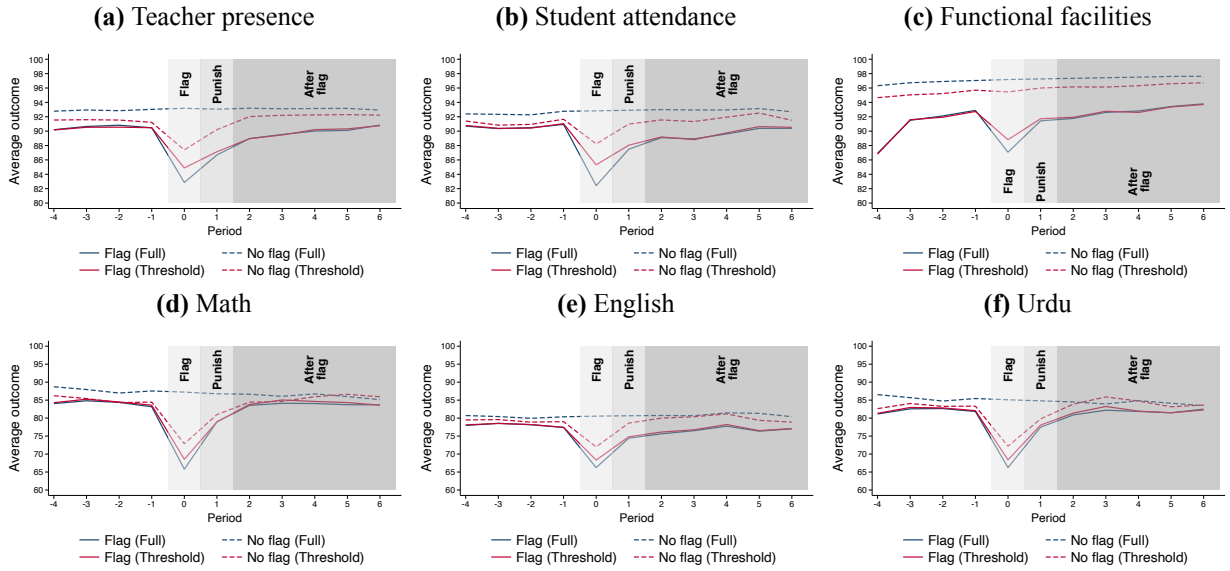
Descriptive trends We report results for the full and threshold samples. For the threshold sample we identify the maraakiz within an optimal bandwidth on either side of the flagging threshold in event-time 0 (Calonico, Cattaneo and Farrell, 2020). We obtain optimal bandwidths separately for each event panel to build a stacked-threshold sample. Figure D1 presents the evolution of the stacked outcomes in relative time. 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*.

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 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. This patterns are similar

[1]Because a bandwidth is selected separately for each outcome and each monthly event, and the running variable is bounded, a linear fit produces bandwidths that vary erratically across events: in some they collapse to near-zero width, in others they span most of the support and no longer isolate units close to the threshold. The design in Appendix C is less exposed to this, since it estimates each event separately and aggregates, so any single poorly behaved window have less weight over the aggregation. In this section, however, a single erratic window in one event changes which units identify the estimates in every period. The quadratic yields better behaved bandwidths.

to an Ashenfelter dip (Ashenfelter, 1978; Ashenfelter and Card, 1985; Heckman and Smith, 1999) where selection into the treatment happens because of a negative shock, and a reversion to the mean might be confounded with an effect.

Figure D1: Evolution of school outcomes in relative time - markaz flagging



Note: The figure presents the average evolution of schools in flagged (continuous) and non-flagged (dashed) 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 are measured as the share of correct answers in standardized exams.

Average effects of flagging Table D1 show the average coefficients of Figure 3 (c). The first column for each variable reports the full sample, and the second the threshold sample. Panel A is for schooling outcomes. 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 for the student test scores show the same pattern. The results imply that the accountability triggered by flagging in the command-and-control 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.

Robustness A concern is that responses might have happened in expectation of the scheme implementation. We test whether officials responded differently when its flag-triggered accountability routines were activated against them for the first time. As such, first-time flagged AEOs should have the highest immediate incentives to avoid punishment. The event study of Figure D2 shows no significant improvement. We also test for fewer post-periods in Figure D3 to test the sensitivity of the results to the data structure. The patterns remain consistent with a reversion to the mean.

We estimate the effects for orange flagging in Figure D4 to validate that no effects are perceived in higher thresholds. Figure D5 test for tehsil-level flagging to explore effects on larger administrative unit. We find no significant effects. We compare maraakiz with the same flagging path before the negative shock to account for selection from maraakiz constantly flagged. We do so by re-building the stacked dataset but allowing flagging in the pre-periods and build a 'flagging history FE' such that we compare maraakiz that experienced exactly the same flagging behaviour. Table D2 shows the results. The coefficients of *After flag* are positive, but, when comparing it against flag *T*, the total effects is small and negative, consistent with our previous results.

We test for alternative estimators. Figure D6 uses Sun and Abraham (2021) estimator, assuming markaz remain 'treated' after the first time flagged. Figure D7 implements the DiD_l estimator by de Chaisemartin and D'Haultfœuille (2026) to account for the turning on/off of the treatment. The results follow the same patterns. Figure D8 show the main event-study using an stacked dataset not imputing the missing data as described in Appendix B.1. The results are similar and the pattern remain consistent with a reversion to the mean for all outcomes. Finally, we estimate in Figure D9 the effect of flagging punishment on the next month performance using a standard regression discontinuity design. There are no significant effects of flagging on the immediate recovery.

D.2 Efficacy of Punishment

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, we estimate equation 4 on first-differenced outcomes over time. Figure D10 plots over time impacts. As positive coefficient indicates a faster recovery path, while a negative indicates the presence of a shock (relative to the period immediately before). We observe a negative shock during the flagging month ($t = 0$). This negative shock is followed by a quick recovery when punishment occurs ($t = 1$). If it were the case that punishment was contributing to an improvement *beyond* the pre-existing recovery path, 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.

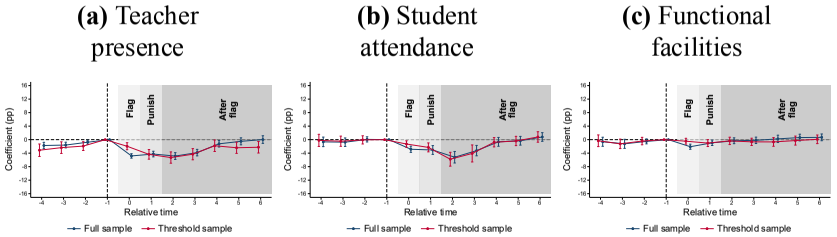
Focusing on the threshold sample, thus analyzing more comparable maraakiz, we find evidence for the efficacy of punishment only in the case of teacher presence. By $t = 2$ the first-differenced outcome is positive and significant. This shows that the rate at which teachers return to school is increased in the first month after flagging by 2 percentage points. However, there are no difference between flagged and non-flagged schools after $t = 2$, suggesting that punishment is not bringing any further improvement. The results for flagging on other outcomes are all indistinguishable from zero. These results show that there is a small impact of command-and-control approaches that only occurs in the short-term on the margin most responsive to hierarchical pressure: personnel attendance. All other dynamics are equivalent to a reversion to the mean.

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	-5.467***	-1.539***	-6.443***	-1.934***	-3.350***	-1.716***
	(0.120)	(0.152)	(0.152)	(0.145)	(0.209)	(0.275)
T×Punish	-2.565***	-1.493***	-3.002***	-1.741***	-1.425***	-1.225***
	(0.141)	(0.204)	(0.161)	(0.247)	(0.169)	(0.291)
T×After flag	-0.449***	-0.624***	-0.887***	-0.771***	-0.403**	-0.860***
	(0.094)	(0.147)	(0.078)	(0.134)	(0.165)	(0.297)
Sample	Full	Threshold	Full	Threshold	Full	Threshold
N. of obs.	8,202,324	673,614	6,080,852	693,441	8,737,364	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	-12.546***	-2.173***	-9.651***	-2.603***	-9.804***	-1.958***
	(0.329)	(0.509)	(0.189)	(0.269)	(0.245)	(0.339)
T×Punish	-2.714***	-0.803	-3.836***	-2.650***	-1.939***	0.062
	(0.413)	(0.723)	(0.229)	(0.380)	(0.308)	(0.470)
T×After flag	-0.503**	-0.231	-1.924***	-1.915***	-0.360*	-0.452
	(0.237)	(0.457)	(0.182)	(0.266)	(0.198)	(0.322)
Sample	Full	Threshold	Full	Threshold	Full	Threshold
N. of obs.	2,750,069	78,564	1,017,391	202,812	2,461,151	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

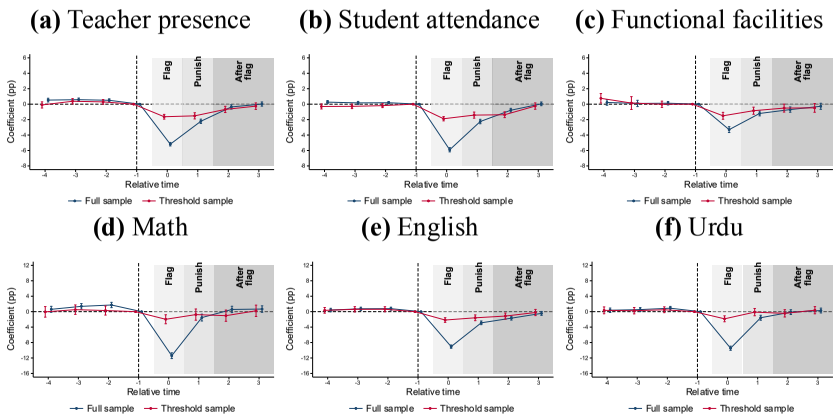
Notes: Results from estimating equation 4. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure D2: 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 D3: 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

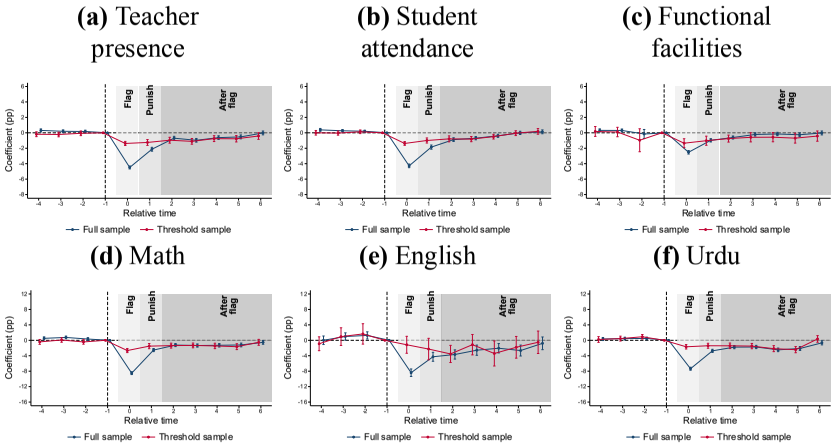
	Dependent variables (range: 0-100)					
	Teacher presence		Student attendance		Functional facilities	
	(1)	(2)	(3)	(4)	(5)	(6)
T	-2.555***	-1.078***	-4.771***	-2.622***	-9.224***	-2.800***
	(0.104)	(0.103)	(0.149)	(0.134)	(0.292)	(0.112)
T×Flag	-3.909***	-1.026***	-2.917***	0.011	1.136***	0.632***
	(0.103)	(0.121)	(0.128)	(0.132)	(0.122)	(0.107)
T×Punish	-0.992***	-0.724***	-0.167	0.043	2.225***	0.337***
	(0.098)	(0.132)	(0.126)	(0.150)	(0.147)	(0.123)
T×After flag	1.460***	0.357***	3.096***	1.278***	4.915***	1.312***
	(0.110)	(0.113)	(0.165)	(0.148)	(0.235)	(0.117)
Sample	Full	Threshold	Full	Threshold	Full	Threshold
N. of obs.	12,160,223	2,332,129	11,281,356	2,946,426	13,858,964	2,440,064
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	-3.647***	-0.995**	-4.123***	-1.875***	-3.617***	-1.200***
	(0.329)	(0.389)	(0.152)	(0.172)	(0.229)	(0.250)
T×Flag	-11.630***	-1.807***	-7.081***	-2.168***	-9.128***	-1.669***
	(0.347)	(0.515)	(0.154)	(0.200)	(0.234)	(0.324)
T×Punish	-1.769***	-0.573	-1.784***	-1.825***	-1.326***	-0.148
	(0.399)	(0.615)	(0.152)	(0.226)	(0.276)	(0.430)
T×After flag	1.722***	0.455	1.590***	0.006	1.822***	0.442
	(0.318)	(0.489)	(0.170)	(0.201)	(0.220)	(0.330)
Sample	Full	Threshold	Full	Threshold	Full	Threshold
N. of obs.	2,871,721	97,336	1,999,108	737,392	2,666,795	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

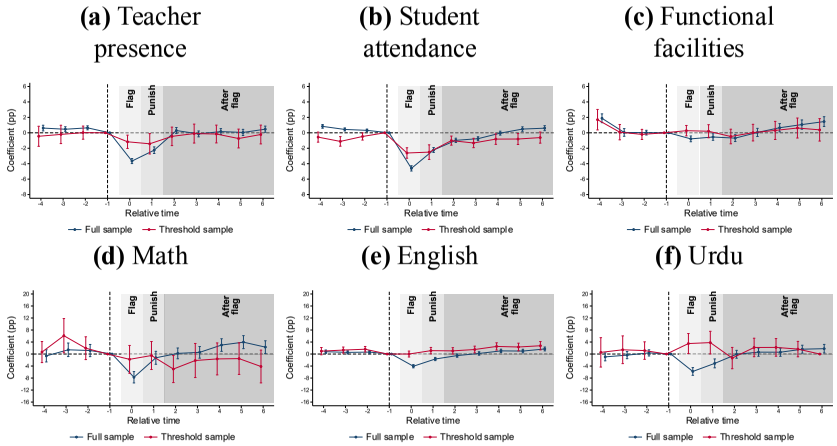
Notes: Results from estimating a modified version of equation 4, including flagging history FE instead of markaz FE. Flagging history is built from concatenating the flagging status in the three periods before the observed flagging. The specification compares maraakiz that had the same flagging path 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 D4: Event study flagging effect on performance orange threshold



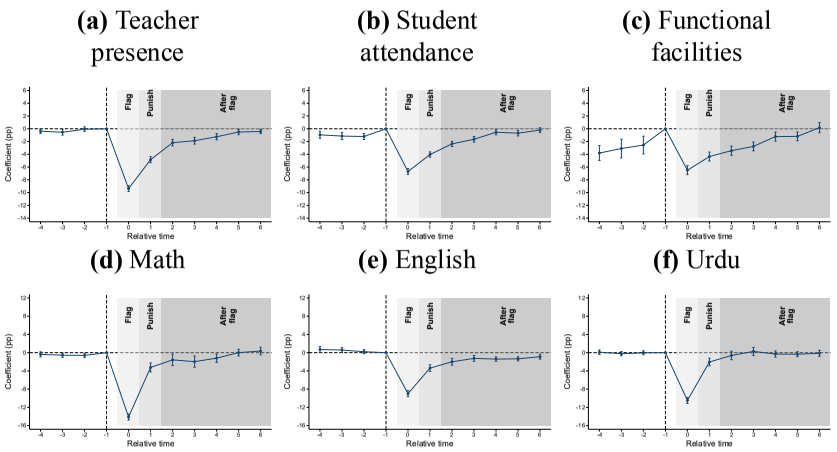
Note: This figure presents results from estimating event studies based on equation 4 using -1 as the base period, comparing schools in orange-flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through 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 D5: Event study - flagging effect on performance tehsil-wing flagging



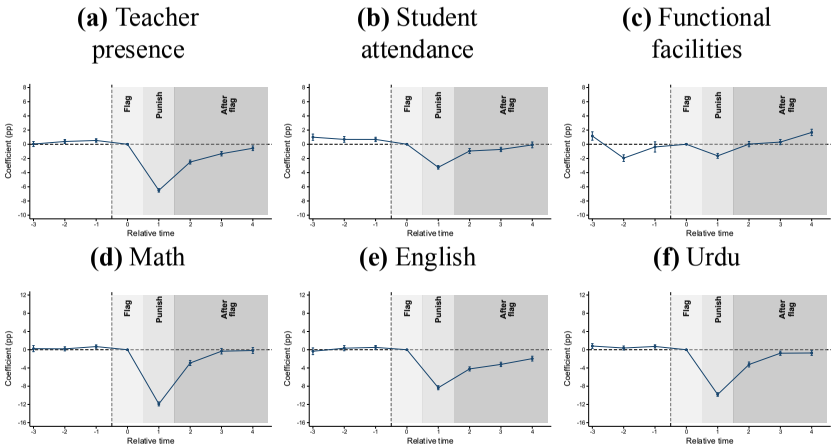
Note: This figure presents results from estimating event studies based on equation 4 using -1 as the base period, comparing schools in flagged and non-flagged tehsil-wing. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through 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 D6: Alternative specifications
Sun and Abraham (2021)



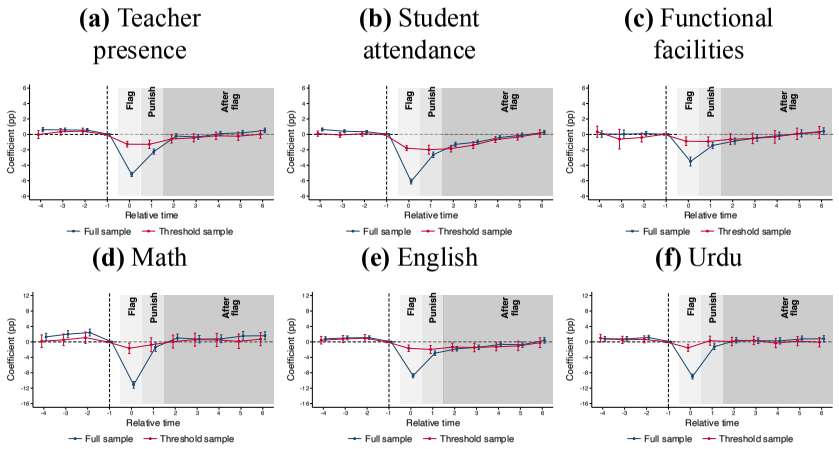
Note: This figure presents the results from estimating an event study based on the Sun and Abraham (2021) difference-in-differences estimator, using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The results are for flagging on the variable in the title of the panel. Teacher 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 D7: Alternative specifications
DID_l de Chaisemartin and D'Haultfœuille (2026)



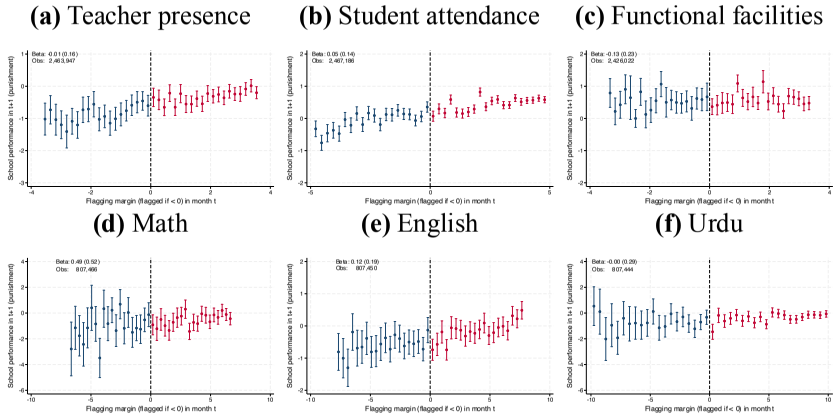
Note: This figure presents the results from estimating an event study based on the DID_l de Chaisemartin and D'Haultfœuille (2026) difference-in-differences estimator, using -1 as the base period, and three placebo periods before the treatment, comparing schools in flagged and non-flagged maraakiz. The 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.

Figure D8: Event study - flagging effect on performance non-imputed data



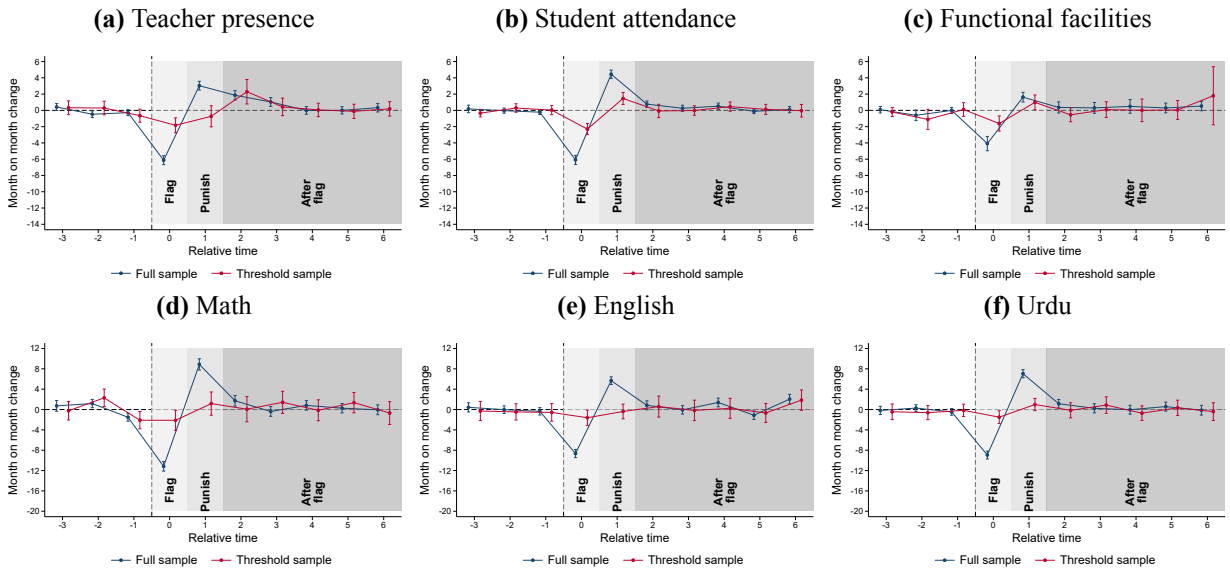
Note: This figure presents results from estimating event studies based on equation 4 using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression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

Figure D9: RDD flagging effect on punishment period



Note: This figure presents the RD plot within the optimal bandwidth (Calonico, Cattaneo and Farrell, 2020) resulting from estimating local linear regression discontinuity specification using as running variable the month t markaz performance (based on which maraakiz are flagged), and as outcome the school-level outcomes for the month $t+1$ when the punishment due to underperformance occurs. We aggregate the school-by-month data at the markaz-by-month level. The coefficient points represent the predicted mean in the bin with its 95% confidence interval. We residualize markaz and month fixed effects as flagging can occur on various occasions over time and across markaz. The left (blue) of the running variable indicates the school is in a markaz that is flagged. The top-left of the figure displays the estimated robust-coefficient and the number of observations. 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 variables are measured as the share of correct answers of students in standardized exams.

Figure D10: 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. The specification compares schools in flagged and non-flagged maraakiz in consecutive months. The blue line is for the full sample. The red line is for the threshold sample. The results are for flagging on the variable in the title of the panel. *Flag* is the period in which information is collected, and the markaz is flagged. *Punish* is when the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting. 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–.

References for Appendix

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.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics*, 60(1): 47–57.

Ashenfelter, Orley, and David Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *The Review of Economics and Statistics*, 67(4): 648–660.

Bank, World. 2017. *World Development Report 2018: Learning to Realize Education's Promise.* Washington, DC: World Bank.

Barber, Michael. 2013. "The Good News from Pakistan." Policy report, accessed May 18, 2026.

Barber, Michael. 2017. "Michael Barber's Views about Punjab's Social Sector (Health & Education)." *YouTube*, accessed May 3, 2026.

Barber, Michael. 2021. *Accomplishment: How to Achieve Ambitious and Challenging Things.* London:Penguin UK. accessed May 18, 2026.

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

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.

Cattaneo, Matias D., Brigham R. Frandsen, and Rocio Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. 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.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2026. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." *The Review of Economics and Statistics*, 1–18.

Education Commission. 2023. "Deliberate Disrupters: Can Delivery Approaches Deliver Better Education Outcomes?" Education Commission DeliverEd Final Report. accessed May 18, 2026.

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.

Khan, Kashmali. 2018. "Punjab, Pakistan: Using Open School Data to Improve Transparency and Accountability." UNESCO International Institute for Educational Planning (IIEP), Paris, France. Open access publication.

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

Malik, Rabea, and Faisal Bari. 2023. "Improving Service Delivery Via Top-Down Data-Driven Accountability: Reform Enactment of the Education Road Map in Pakistan." Education Commission and Blavatnik School of Government DeliverEd Initiative Working Paper. accessed May 18, 2026.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199. Themed Issue: Treatment Effect 1.