Designing and Implementing Experiments within Local Bureaucratic Systems: A Cautionary Tale from an Educator Incentive Program†

Saher Asad[a], Masooma Habib[b], Naureen Karachiwalla[c], Katrina Kosec[d], Clare Leaver[e], Attique ur Rehman[f]

May 24, 2024

Partnering with governments to co-design pilot interventions and embed them in local bureaucratic systems is increasingly seen as 'best practice' on grounds of scalability and sustainability. This paper reports on a pilot program that was co-designed with, and embedded within, the Elementary and Secondary Education (E&SE) Department in Pakistan's Khyber Pakhtunkhwa province that offers a cautionary tale. The over-arching desire to work within existing bureaucratic systems, while laudable, constrained the design of the randomized controlled trial. This paper presents findings on some of the institutional factors which resulted in failed implementation of the RCT and a lost opportunity to learn about the efficacy of key design features. The paper briefly outlines the design of the pilot—promotion-based incentives for educators—and summarises the largely null results. It then turns to implementation, discussing what went wrong, how this was uncovered, and lessons learned for co-designing and embedding future pilot studies with(in) government.

JEL Codes: I28, I25, H83

Keywords: educational economics, teacher policies, teacher incentives

a. Saher Asad, World Bank, 1818 H St NW, Washington DC, 20433. saher.asad@gmail.com.
b. Masooma Habib, Consortium for Development Policy Research (CDPR), 19 A FCC Scheme, Maratab Ali Road, Gulberg IV, Lahore, Punjab, Pakistan. mhabib@gwmail.gwu.edu.
c. Naureen Karachiwalla, International Food Policy Research Institute (IFPRI), 1201 I Street NW, Washington DC, 20005. n.karachiwalla@cgiar.org.
d. Katrina Kosec (corresponding author), International Food Policy Research Institute (IFPRI), 1201 I Street NW, Washington DC, 20005. k.kosec@cgiar.org.
e. Clare Leaver, Blavatnik School of Government, University of Oxford, 120 Walton St, Oxford OX2 6GG, United Kingdom. clare.leaver@bsg.ox.ac.uk.
f. Attique Ur Rehman, University of Delaware, 303 Alfred Lerner Hall Newark, DE 19716. attique@udel.edu.

† We thank Barbara Bruns, David K. Evans, Derek Neal, Pieter Serneels, Andrew Zeitlin, and audiences at IFPRI, the University of Sherbrooke, and the Center for the Study of African Economies (CSAE) for helpful comments and feedback. We acknowledge excellent research assistance from Charles Gale, Amir Jilani, and Umair Kiani, as well as excellent survey management from Research Consultants (RCons). We also thank Jeffrey Patton for invaluable advice regarding test development and analysis. This work was undertaken as part of the CGIAR Research Initiative on Fragility, Conflict, and Migration and the CGIAR Research Initiative on National Policies and Strategies. We would like to thank all funders who supported this research through their contributions to the CGIAR Trust Fund: https://www.cgiar.org/funders/. It also benefited from support of the International Growth Centre, the Lahore University of Management Sciences, the University of Oxford, and the U.S. Agency for International Development (USAID) [cooperative agreement # AID-391-IO-11-00002]. The Center for Development Policy Research (CDPR) provided logistical support in Pakistan. The authors do not have any financial conflicts of interest to disclose.

1 Introduction

Scalability and sustainability are increasingly recognized as essential features of interventions to improve government service delivery, but they are often elusive (Bold et al. 2018; Ganimian 2020; Araujo et al. 2021; List et al. 2021; List 2022). Many have posited that partnering with government to co-design pilot programs, embed them in local bureaucratic systems, and nurture relationships over time can help (Carter et al. 2021; Pappas 2021). Doing so serves to: ensure government buy-in; overcome information asymmetries (by tapping into government's often superior information about service delivery problems and potential solutions); promote greater policy coherence; and avoid costly transitions from researcher or Non-Governmental Organization (NGO) implementation to government implementation. On the other hand, such partnerships can sometimes limit researchers' options, as government may require certain aspects of intervention design or targeting, and researchers must operate within current institutions. In theory, this could result in a lower intensity of implementation, spillovers, or even enable "gaming." This paper explores these issues, drawing on the experience of a pilot reform instigated in 2017 by the Elementary and Secondary Education (E&SE) Department in Khyber Pakhtunkhwa (KP) province, Pakistan.

The E&SE Department was worried about the state of learning in KP. A previous review paper, commissioned by the Department and conducted by a member of our research team, had noted that strengthening the monitoring and rewards system for educators could benefit student learning (Habib 2015). Reflecting on this review, the Department approached our research team to co-design a randomized control trial (RCT) to evaluate a new promotion policy for teachers and head teachers. While performance assessments already existed, they were generally unstructured and did not measure factors known to promote student learning. The reform sought to collect objective and informative performance measures for determining

teacher and head teacher promotions, with the aim of improving educator effort, teaching practices and, ultimately, student outcomes.

The project faced various implementation challenges driven by local institutional features and yielded largely null impacts. It did, however, generate rich data, including video evidence, on how implementation unfolded. We use these data to document what went wrong, and to draw out lessons for future studies seeking to co-design interventions with government and embed them in local bureaucratic systems.

The pilot implemented two interventions during the 2017-2018 school year in rural public primary schools in three administrative districts. District education officials visited schools and evaluated teachers and head teachers on several aspects of their work including their and their students' attendance, teacher pedagogy, and student learning. Data were recorded on tablets, with photographs certifying attendance and videos of classroom observations and student performance on an oral math assessment. In both treatment arms, the relevant educators (teachers or head teachers) were informed that their scores would be compared to others in the same school district, and relative performance within their school district would determine the speed of their promotions. The E&SE officials were trained on both the interventions, and on RCTs and the importance of keeping the experimental groups separate to ensure the integrity of the study. Importantly for what follows, in each school district, at the Department's request, the same E&SE official conducted the interventions in the eight treatment schools as well as the four control schools where their previous practices for school visits were to be continued, which typically included providing advice to teachers.

We find null effects of treatment on all primary outcomes specified in our trial pre-registration:[1] student test scores in math, attendance, grade promotion, teacher pedagogical

[1] https://www.socialscienceregistry.org/trials/2815.

practices, educator attendance, and educator effort. Rather, some outcomes (test scores and teacher pedagogy) improved between baseline and endline, either for *all three* or for *two* of the experimental groups. In particular, there was an economically meaningful improvement in teacher pedagogy in both treatment *and* control.

Concluding our analysis with this finding would result in a lost opportunity to learn from the experiment. Given the design of the pilot, we have documentary evidence of what happened via videos recorded as part of the pilot in treatment schools. These videos provide novel information on what went wrong and why. Research assistants watched the videos (of teachers teaching grade 4 math and students taking the oral math exam). These recordings show that, rather than switching to the role of passive performance monitor in treatment schools as intended, many E&SE officials continued to engage in the active coaching practices that they had used prior to the study. For example, in ¼ of treatment schools, we saw officials, who should have been monitoring, coaching teachers to split students into groups and use more learning materials (pedagogical practices that scored highly on the classroom observation rubric). Video recordings of the administration of oral math test that was also part of the pilot also show officials providing hints and sometimes answers to students.

We also conducted eight semi-structured interviews with E&SE officials, representing 40 percent of the 20 officials in the study, which helped to understand behavior in control schools, where we lack videos.[2] Corroborating our video evidence, seven of the eight officials gave examples of their coaching of teachers in treatment schools, during and after pedagogy observations, and explained that they wanted to continue to give feedback. Moreover, seven of the eight spoke openly about encouraging *control group* teachers to adopt good teaching practices, and/or warning them about a forthcoming accountability system that would

[2] Unfortunately, we could not reach all 20 officials because some had transferred to other jobs.

incentivize such practices. The design of the classroom observation tool that district officials used in the pilot may have focused them on these practices, which they then brought to their business-as-usual coaching in control schools. Consistent with this, we see that teaching practices, such as having students work in groups rather than rote teaching, improve in all three groups (which may also explain the improvements in test scores in all three groups).

Other factors also affected the pilot program rollout. The Memorandum of Understanding (MOU) between the research team and the E&SE Department was initially broad, signed with a supportive Education Director who was later replaced, along with the civil servant championing our project. Although we suggested evaluating and rewarding officials based on the quality of school visits for incentive compatibility, the E&SE Department deemed it inappropriate at their level.

In many trials with null results, researchers must conjecture underlying mechanisms. Here, with the help of the videos and interviews with E&SE officials, it is unambiguous that the pilot program was not implemented as intended. Reflecting on this evidence, we identify four lessons for co-designing and embedding future pilot studies with(in) government. First, turnover is a big problem in the public sector, raising the importance of a detailed MOU that will preserve institutional memory and commitment to design *and* intended implementation details. Second, qualitative work conducted prior to designing a trial can elucidate what norms and practices a proposed intervention might cut across, and steps that could be taken in mitigation. Third, caution is needed before asking a civil servant to take on different roles requiring different "mental models," particularly if this means asking an official to *stop* performing a function that they think they should be doing. Finally, training for government implementers may need to be more extensive and continuous, involving various check-ins and reminders to stick to protocols and minimize reversion to existing practice.

Discussions of how to get implementation right have not been at the forefront of the literature on education systems. To date, few papers have discussed implementation failures, or how to tackle them. Banerjee et al. (2017) reviews challenges that can arise when moving from "proof of concept" interventions run by NGOs to government implementation at scale. One of the papers, Bold et al. (2018), employs an RCT to compare the effectiveness of a (small-scale) contract teacher intervention. It assesses outcomes when implemented by an NGO versus the Kenyan government. Replicating earlier work, the authors find a sizeable positive treatment effect on student learning for the NGO-implemented intervention, but no effects of the government-implemented intervention. Unfortunately, the authors lack data to draw more than tentative conclusions as to why the government treatment arm failed (they hypothesize it may have been political resistance and low monitoring capacity of government). Our paper contributes to this literature by demonstrating the value of data on sources of implementation failure.[3] As it becomes more widely accepted that null results should be reported, collecting data to document *why* interventions failed, and how much was due to improper implementation, will become increasingly important.

The remainder of the paper is organized as follows. Section 2 describes the interventions. Section 3 presents our sampling and empirical strategy. Section 4 characterizes our (broadly null) results. Section 5 discusses what we think explains these results. Finally, Section 6 concludes with lessons on how to capitalize on the benefits of co-designing interventions with government while avoiding costs.

[3] See also recent work by, e.g., Shaffneret al (2021) and Muralidharan and Singh (2023), who take a mixed methods approach to uncover reasons behind null results.

2 Intervention

2.1 Existing Policies

Teachers in KP are trained under standard E&SE Department programs in the Provincial Institute of Teacher Education. Pre- and in-service training includes preparation of lesson plans to cover relevant curriculum at grade-appropriate levels, the use of learning materials, and how to carry out child-friendly approaches to teaching. Teachers are taught that fourth grade mathematics classes must follow pre-developed lesson plans in consultation with head teachers and district officers. Multi-grade classrooms are the norm in primary schools, often due to teacher shortages or fewer student numbers in higher grades. Students are often grouped according to their grade rather than their ability despite evidence that 'teaching at the right level' where students are grouped together by ability) substantially improves student outcomes (Banerjee et al. 2017). Further, schools are traditionally comprised of a single gender of student.

The KP government has accountability policies in place intended to lead to high effort from educators. It was within these structures that we tried to design and implement an intervention jointly with the government. All civil servants are evaluated annually using the Performance Evaluation Report (PER), which should carry rewards and consequences for promotions and motivate high effort from educators (Pakistan Public Administration Research Centre, 2004). In practice, PERs are largely ineffective for this goal.

The PER system for teachers suffers from several problems. First, there is no teacher-specific PER, only a general civil servant PER that does not directly measure the quality of teaching and is highly subjective (Tanwir 2010).[4] As Holmstrom and Milgrom (1991) show,

[4] The PER includes measurements such as personal qualities (for example, perceived integrity), and the guidelines even stipulate the type of typewriter that should be used to write the report.

employees may focus effort on tasks that are measured and rewarded, meaning that the present PER system may not incentivize high-quality teaching. Second, PERs are conducted at the end of the calendar year, covering half of two school years, making it unclear which teacher deserves the credit or blame for student outcomes. Finally, PERs are carried out by the head teacher, but head teachers are not necessarily independent evaluators, and most teachers tend to receive the same, passing score (i.e., ratings compression) (Pakistan Public Administration Research Centre, 2004), rendering PERs uninformative. Finally, PERs lack clear performance targets, and there are not provisions for sanctions or punitive measures if teachers perform poorly (Tanwir and Chaudhury, 2016).

Under the PER system for head teachers, an E&SE official called an Assistant Sub-District Education Officer (ASDEO) visits the school (this is called a school inspection). The official policy is that these officials should carry out inspections quarterly and issue warnings to head teachers not complying with policies and regulations.[5] In practice, inspections are irregular, there is little guidance on what they should entail, inspectors observe and note schooling aspects that are not directly meaningful for learning (e.g., students' cleanliness), and warnings to higher-ups are rarely issued. Inspections thus carry little meaning. This is unfortunate, as a study from India (Muralidharan et al. 2017) found that monitoring schools —even without financial rewards—had one of the strongest impacts on teacher attendance compared to other inputs (including hiring more teachers).

Despite the existence of these two accountability systems, promotions are largely based on seniority, academic qualifications, and connections, providing no *de facto* incentives. The civil service promotion document explicitly enumerates the requisite trainings, number of years of service, and educational qualifications required to be promoted to a certain level.

[5] Again, these regulations do not pertain to learning but rather, to more general civil service rules.

Consequently, newer teachers who are hard-working and creative in their teaching do not get rewarded for their effort or for producing better learning outcomes among their students.

2.2 Pilot program

The E&SE Department co-developed an RCT with our research team following discussions about policies that might improve learning outcomes (Habib 2015). It sought to strengthen teacher and head teacher career incentives by reforming evaluation and promotion processes for a randomly selected 160 treatment schools out of 240 study schools. It comprised two interventions (each brought to 80 schools) built on theory about best practices to incentivize educators (Barlevy and Neal, 2012; Steinberg and Donaldson, 2016; Karachiwalla and Park, 2017; Cilliers et al. 2018; Gilligan et al. 2019). For both, the E&SE Department chose to enlist ASDEOs—the E&SE officials traditionally responsible for school visits (many of whom were previously teachers). The ASDEO was responsible for visiting all schools in their school district (which are separated by gender) and none outside, as this is how workloads are officially divided. Prior to hearing about the interventions, ASDEOs were asked to jointly prepare a "business as usual" protocol describing current visit practices. These protocols point to active engagement and coaching with educators and children (see Appendix C). ASDEOs then received training on the new monitoring role to be used in the 160 treatment schools and were asked to continue "business as usual" in the 80 control schools.[6] They were also trained on RCTs and the importance of keeping experimental groups distinct.

Eighty schools assigned to the *teacher treatment* (T) were to receive three unannounced visits (randomly assigned by the research team) collecting data on teaching-specific outcomes relevant to learning: presence of the grade 4 teacher and of their students,

[6] This training occurred in October 2017 and lasted three days, involving ASDEOs in addition to higher-level E&SE officials who would not deliver the trainings but would oversee ASDEOs.

the grade 4 math teacher's pedagogy (measured using a Stallings-type classroom observation tool (World Bank 2015)), and test scores of the teacher's students on a researcher-designed oral math test (test construction is described in Appendix Section B.1). Another 80 schools assigned to the *head teacher treatment* (HT) were also to receive three unannounced visits, with information collected on presence of the head teacher, all teaching staff, and all students in the school, and the pedagogy of two randomly selected teachers (using the same tool).

In both treatments, scores over all visits would be aggregated into a single, cardinal score for each educator and linked to career progression via promotion tournaments within school districts; the teacher (respectively, head teacher) who performed the best relative to others in the same school district and treatment arm would have their promotion fast tracked by a year compared to their expected time of promotion, while the teacher (head teacher) who performed worst would have it delayed a year compared to their expected time of promotion (the two other teachers were unaffected). In control schools, there was no change in the promotion system. As the number of educators whose promotions would be moved forward by one year was equal to the number of educators whose promotions would be delayed by one year, these interventions were budget neutral. Note that we only selected schools where a teacher or headteacher was at least two years from their next expected promotion, so as not to delay promotions that were already counted upon.

Teachers and head teachers in treatment schools were informed of the new performance incentive program during the ASDEO's first visit to the school. The program was communicated to all educators as the "Improving Khyber Pakhtunkhwa's Accountability in Schools (IKAS) project." Grade 4 educators received detailed scoring rubrics (described in Appendix Section B.2) and were instructed to tell nothing to educators in the control group. Each of the four metrics collected in treatment schools were recorded on tablets using photos

(for attendance)[7] or videos (for classroom observations of pedagogy and oral math assessments of students). Tablets required a school ID code to access the system, intended to prevent use in control schools. To prevent manipulation of scores without enhancing learning, some details of the scoring process were withheld from educators (e.g., which pedagogical activities earned more points, and math test questions)—though it is possible that some information leaked.

ASDEOs conducted both types of visits, receiving information on the school to visit and its treatment status the evening before the planned visit, and were instructed not to notify the school. If they could not visit that school that day, they had to provide a reason, and were assigned a randomly selected back-up school if the reason was school-specific (e.g., school closure) and not, for example, due to illness. ASDEOs received both treatment protocols; a tablet (pre-loaded with a computer assisted personal interview (CAPI) program that included teacher and student rosters from the research team's baseline survey); a tripod stand for video recordings; and a Wi-Fi box with credit for data uploads.

To incentivize ASDEOs to undertake reformed inspections thoroughly and objectively, they were notified that scores would be audited by individuals with an arms-length relationship to the ASDEO/school: either a district official from a non-study administrative district or a secondary school teacher from an administrative district in the study. Auditors were to review information collected during visits (photographic proof of attendance, video classroom observations) and note several aspects of the visit (e.g., if data were missing, if the video camera's angle captured the teacher and all students, if photos of teachers—who were asked to hold their IDs—were taken correctly, and if any absences were excused for valid,

[7] In both treatments, educator attendance was captured by a photo of the teacher holding their ID card or a photo of a letter of an excused absence. Unexcused absences were noted, requiring no photo.

properly-documented reasons). ASDEOs performing the visits were also to have their promotions moved forward, maintained, or delayed based on the audits.

3 Study Design and Empirical Strategy

3.1 Sampling

With the E&SE department, we identified three administrative districts in which it would be both logistically and politically feasible to operate in rural, public primary schools: Charsadda, Mardan, and Nowshera. We compiled a list of all rural, public primary schools and eliminated those where class sizes were too big to manage, or teachers or head teachers were within two years of their next expected promotion. We needed school districts with at least 12 eligible schools; 20 met this criterion, yielding 240 schools.[8] In each school district, four schools were randomly assigned to each of our three study arms. Figure A1 displays the project timeline. Details on sampling and randomization are in Appendix Section B.3. Appendix Section B.4 describes our power calculations. We conducted a baseline survey shortly after the beginning of the school year (September – October 2017) and an endline survey shortly before its end (February – March 2018), resulting in a panel of students and educators.

3.2 Empirical Strategy

We begin by examining changes in our pre-registered outcomes between baseline and endline. We report paired *t*-tests for differences between baseline and endline means for each of the three study arms in graphical format. Student-level outcomes are: math scores from

[8] Note that the average distance to the nearest school of a different treatment status is 2.6 km. While our preference was randomization at the school district level, few met the inclusion criteria. We thus aimed to reduce spillovers via intense training on RCTs and providing a letter from the education department (note that this was not an official directive but a letter indicating support of the study.

the survey-based endline test (proportion correct out of 45 questions);[9] whether the student was present on the day of the endline survey; and whether the student dropped out between the baseline and endline. Educator outcomes include teaching practices and effort. In the first category we include: whether the teacher helps students individually outside of school; whether the teacher splits students into groups for learning activities;[10] and whether the teacher has spoken to the student's parents regarding their progress in school.[11] In the effort category we include whether the grade 4 teacher and the head teacher are present on the day of the endline survey; and whether the head teacher helps with the grade 4 math class.[12]

We then estimate reduced form equations to measure intent to treat effects:

$$Y_{it} = \beta_0 + \beta_1 T_i + \beta_2 HT_i + Y_{i,t-1} + \gamma_c + \varepsilon_{it} \qquad (1)$$

where Y_{it} is an outcome for unit i (a student or educator) at time t. T_i and HT_i are indicators for schools assigned to the teacher treatment and head teacher treatment, respectively. $Y_{i,t-1}$ is the baseline value of the outcome (McKenzie 2012). School district fixed effects, γ_c, are included as randomization was stratified by school district. Coefficients of interest are β_1 and

[9] This written test was part of the baseline and endline surveys (tests differed slightly for each round) to measure study outcomes and differed from the oral math test of the pilot intervention.
[10] This and the previous question were, respectively: "Does your math teacher, [name of grade 4 Math teacher], give you extra help in math individually either during school or after school (either if you have asked or if you have not asked for extra help)?"; and "Does your math teacher, [name of grade 4 Math teacher], put students into groups to work on math problems?" The options for both were: 1 – Every math lesson, 2 – Most math lessons, 3 – Some math lessons, 4 – A few math lessons 5 – Never. We coded an indicator equal to one if the response was every or most math lessons, and zero otherwise.
[11] Students were asked, "Has [name of grade 4 Math teacher] given feedback to your parents about your progress in math this school year (yes or no)?"
[12] Since this grade and subject was the focus of the study, this proxies for whether the head teacher extended further effort given the incentive to do so.

β_2, the impacts of the teacher and head teacher treatments, respectively. Standard errors are clustered at the school level, the level at which treatment was assigned (Abadie et al. 2017).[13]

4 Results

4.1 Balance and attrition

Appendix Table A1 presents student-level summary statistics (columns 1, 3, and 5) and number of observations (columns 2, 4, and 6) and p-values from tests for balance across pairs of study arms (columns 7-9). Appendix Table A2 shows the same for school level outcomes. In both cases, we find strong balance across study arms. Further, there is no differential attrition by treatment group for students (Table A3, Column 1) or teachers (Column 2), though head teachers in the teacher treatment arm are more likely to attrit (Column 3).[14]

4.2 Implementation

Table 1 shows statistics on implementation of the pilot program. In teacher treatment schools, 81 percent of teachers reported "far more" visits than usual, and an additional 11 percent reported "more" visits than usual, while these numbers were a similarly high 82 percent and 8 percent among head teachers in head teacher treatment schools.

In teacher treatment schools, 93 percent of teachers had heard about the IKAS project specifically, while this number was 95 percent among head teachers in head teacher treatment schools, and 30 percent in control schools among both teachers and head teachers, indicating that educators in many control schools knew about the program. Note that we

[13] We correct for multiple hypothesis testing within each family of outcomes (student outcomes, teacher practices, and educator effort) using the family-wise error rate described by Anderson (2008), and report adjusted p-values.

[14] One school had merged with a non-sample school by the time of the endline survey, so we did not visit that school at endline, and our final sample comprises 239 schools.

asked about the specific name of this program. Still more control group educators may have heard about the program, but not recognized its name and thus indicated they had not heard about it—making 30 percent a lower bound on information spillovers about the program.

During the 2017-2018 school year, an average of 2.7 visits (2.6 visits) were completed in the teacher (head teacher) treatment arm, both lasting 7 hours on average. Some schools received only two of the intended three visits, while others received all three. We did not measure the reported length of visits in control schools, but the "business as usual" protocol completed by ASDEOs during training (prior to learning about the pilot) indicated visits of only 50 minutes, on average, with between 3 and 6 visits per year.

Educators also appear to have understood what ASDEOs were measuring and how it would affect them. Among teachers (respectively, head teachers) in the teacher (head teacher) treatment arm, 66 percent (60 percent) correctly said the new visits measured their own attendance, 76 percent (61 percent) correctly said they measured student attendance, 70 percent (54 percent) correctly said they measured teacher pedagogy, and 79 percent (67 percent) correctly said that good performance on the new evaluation system would increase their chances of promotion. Further, among teachers in the teacher treatment arm, 56 percent correctly said the new evaluations measured student learning. (Student learning was *not* measured in the head teacher treatment arm schools, though 47 percent of head teachers erroneously said student learning was measured). For comparison purposes, only 23 percent of incentivised teachers in the teacher treatment arm, and 25 percent of incentivised head teachers in the head teacher treatment arm, incorrectly reported their salary would be increased (but without promotion) due to good performance on the new evaluations.

4.3 Student outcomes

Here we report results for the outcomes described in Section 3.2. Figure 1 presents means of each variable at baseline and endline, and *p*-values from paired *t*-tests for differences between baseline and endline within each of the three study arms. We see statistically significant improvements in test scores for both treatment groups and a similarly sized, but statistically insignificant, improvement for the control group. However, the magnitude of these changes is only approximately one more question correct out of 45.[15] There is a statistically significant improvement in student attendance for the control group and statistically insignificant improvements for both treatment groups, but none of these is economically meaningful. There are statistically significant declines in drop-out for the teacher treatment and control groups, and a statistically insignificant decline in drop-out for the head teacher treatment group, but again, effects are small. Estimating Equation (1), we similarly find no significant treatment effects on any student outcome (Table 2, Panel A). This null result echoes Barrera-Osorio and Raju (2017), who report that an incentive program in Pakistan tying educator bonuses to school enrollment, student exam scores, and exam participation failed to improve exam scores—a finding the authors speculate is due to lack of educator "know-how" of good teaching practices. Considering their results and those from this paper, one might conjecture that there could be positive effects on learning from an intervention that includes both accountability and coaching.

[15] Between baseline and endline, the proportion correct went from 47 to 49 percent for control, from 47 to 48 percent for the teacher treatment, and from 48 to 49 percent, which amount to less than one additional question correct. for the headteacher treatment.

4.4 Educator outcomes

In Figure 2, we consider the six educator-level outcomes described in Section 3.2. There are statistically significant and economically meaningful improvements over time for all three study arms for three of the six: use of student groups (Panel A), reporting progress to parents (Panel B), and the head teacher supporting the grade 4 class (Panel F). These improvements are non-trivial: for all three study arms, we observe a 15-20 percentage point increase in the likelihood of a teacher splitting the class into groups, a 5-10 percentage point increase in the likelihood that parents receive feedback on student progress in math, and a 15-20 percentage point increase in the likelihood that the head teacher helps teach the grade 4 math class.

Turning to the remaining outcomes, for helping with classwork outside of school (Panel C) and teacher attendance (Panel D), there are no economically meaningful changes between baseline and endline for any study arm.[16] For head teacher attendance (Panel E), we observe differently signed time trends (positive for the teacher treatment and control, negative for the head teacher treatment), though none is statistically significant. Estimating Equation (1) for these six outcomes, we again find little evidence of meaningful treatment effects (see Table 2, Panel B).

5 Discussion

In this section, we discuss the challenges that arose during implementation of the pilot program, focusing on how we uncovered them and why they may have arisen. We draw on this discussion when concluding with a set of lessons learned for future work in Section 5.

[16] There is a (weakly) statistically significant improvement in helping with outside classwork in the control group, but the magnitude is trivial.

5.1 Coaching of teachers in treatment schools

The provincial government insisted on using ASDEOs to conduct all school visits. This decision was taken to embed the new accountability system within existing administrative structures. While a laudable aim, this meant that the same individuals were conducting visits in both treatment and control schools. Under "business as usual" visits (per the protocol that ASDEOs created at the start of the training session), ASDEOs actively provided guidance and feedback to teachers. During the pilot, they were asked to continue this practice in control schools but switch to a passive monitoring role in treatment schools. Two sources of evidence suggest they found the change difficult and continued to actively coach teachers.

The first piece of evidence comes from the videos recorded during classroom observations in treatment schools. Appendix Table A4 summarizes the number of visits that occurred, and the number of videos that were available to watch (recall that two classroom observation videos were to be recorded in head teacher treatment schools, but only one in teacher treatment schools). Any missing videos are due to video files being corrupted when uploading or downloading rather than videos not having been recorded.[17]

Research assistants watched each video (for the first and last visit to each school) and noted whether the ASDEO took on an active role rather than the mandated, passive monitoring role in the classroom.[18] For example, active roles could take the form of the ASDEO telling the teacher to do a specific activity (e.g., one known to deliver a high pedagogy score, like putting students into small groups), telling the teacher what language to use, a

[17] The number of videos variable is set to zero in the event of missing videos. Doing so provides a lower bound on the extent of cheating since the implicit assumption is that no cheating has occurred in schools without a video.

[18] Three RAs graded the videos of the classroom observations and the administration of the oral math test over the span of one month. Each was given the same observation tool as the district officials used to score the observations and filled them out as they watched. Each video was watched and scored twice by different RAs.

teacher saying on camera that a district official had previously asked them to do something during this classroom observation (suggesting a conversation in advance of the observation which was not supposed to have occurred), or an ASDEO telling a student what to do.

Table 3, Panel A describes the incidence of the behaviors uncovered while watching the videos. Overall, active coaching-type behaviors were recorded in 19 percent of schools. ASDEOs told teachers what activity to do in 18 percent of schools and in 1.25 percent of schools, told the teacher to teach in English. These behaviors were caught by noting the inspector directly telling a teacher what to do or making eye contact or gestures that elicited a teacher response. Active coaching-type behavior occurred in 23 percent of head teacher treatment schools and 14 percent of teacher treatment schools. Additionally, active coaching-type behavior decreased between the first and last visits (from 11 to 7 percent), but the difference is also not statistically significant (p-value = 0.24).[19] Clearly, coaching may also have occurred off-camera (or was difficult to assess and code), so these already large numbers are likely substantial underestimates.

We also conducted detailed, semi-structured qualitative interviews with eight of the 20 ASDEOs.[20] Transcripts were coded into seven themes (see Appendix D), identified by the prevalence of a topic arising.[21] Two members of the research team coded the interviews separately, cross-checking to ensure consistency.

[19] The greater prevalence in head teacher treatment schools is mechanical as (roughly) twice as many videos were recorded there – though this difference is not statistically significant (p-value = 0.11).
[20] Officials were stratified by experience and whether (based on our video evidence) they coached or not, and then randomized into a sequence to be contacted. There were only 4 female officials in the sample (given a low prevalence of girls' schools in these areas) and 3 of the 4 were interviewed.
[21] The themes are: 1) officials liked the intervention (8 of 8); 2) officials told control schools about the program or implemented the program there (7 of 8); 3) officials felt that teachers made an effort and schools improved because teachers were more motivated (8 of 8); 4) officials coached teachers (7 of 8); 5); teachers objected to the videos at first (5 of 8); 6) the project informed the government's subsequent inspection program (4 of 8); and 7) Other (themes mentioned only by one or two officials).

Echoing the videos, the interviews indicate that officials continued to coach teachers on their pedagogy. One female ASDEO stated, *"after the observation… when the teacher would come to me, and the head teacher would also especially come and ask how did my teacher do her lesson … I would tell her if there was a shortcoming in the lesson, that you did not involve the students, or anything else."* One male inspector also remarked that they would show the teacher the video after the observation.

A possible explanation is that some ASDEOs found it difficult to change their "mental model," and continued to feel responsible for providing help and advice. The "business as usual" protocols make clear that ASDEOs were accustomed to approaching schools collaboratively, and advising the teacher in a kind, supportive manner. Moreover, when asked to reflect on the pilot after the intervention, seven of the eight officials expressed a desire for space to coach, as many were teachers in the past. Comments here included: "*The most important role of the supervisor is to supervise the teaching, to see what the deficiencies in their teaching are and to point them out... [and] if they see something really good in one teacher, then they could convey that to the other teachers"; "We felt sometimes that we should say something to the teacher during the observation"*; and *"in the [other] project that we are doing now, we have the option of giving feedback."*

To further explore why some ASDEOs continued to coach teachers in treatment schools, we looked for evidence of statistical associations between the incidence of such behavior and school, classroom, and ASDEO characteristics. The first column in Table 4 reports results from a school-level regression of the proportion of videos in which a coaching behavior was detected on baseline school and classroom characteristics. The only statistically significant association is a negative one between the proportion of videos in which coaching was detected and the distance of the school from the District Education Office (indicating

that the school is more rural since DEO offices are located in towns). The coefficient is extremely small, however, so we do not take this as strong evidence of inspectors coaching more in more rural schools. The first column in Table 5 reports results from a similar exercise where we regress the coaching indicator variable on ASDEO characteristics, again collected at baseline. Here, the only statistically significant association is for gender: female ASDEOs are more likely to coach teachers during classroom observations than male teachers.

5.2 Coaching during student math tests in treatment schools

We also examined inspector coaching behavior during the administration of the oral grade 4 math test administered in schools in the teacher treatment arm (no such test occurred in head teacher treatment schools). Table A4 shows the size of this sample in the bottom panel, listing the number of visits and the number of oral test videos that were available to watch.

Research assistants watched 104 videos of the district officials conducting oral math tests during the first and last visits to each school. These videos reveal a further implementation challenge. Inspectors deviated from the protocol for administration of the oral math test in several ways. There were three types of student coaching detected: giving a student the answer to a question; hinting at the answer to a question; and (against protocols) allowing the teacher or head teacher to be present during the assessment. The second behavior can be seen as coaching students; the inspector may have wanted to help the students come to the right answer. The third was specifically discouraged in the pilot protocol but, speculating, inspectors may have wanted teachers to learn something from the exercise to improve their teaching—for example, which students were struggling, how to teach a particular concept, etc. Table 3, Panel B shows that at least one of these three behaviors was detected in 36 percent of (teacher treatment) schools. Hinting was the most common (36 percent), with a district official only actually giving a student the answer in 8 percent of

schools. When district officials did offer hints, this help could be quite extensive. For example, one district official hinted at answers in 13 (of 50) questions during one visit. Both hinting at, and giving, answers increased slightly over time, albeit not statistically significantly: district officials hinted in 21 percent of schools during the first visit and 26 percent during the last visit (p-value = 0.37); and gave students answers in 4 percent of schools in the first visit and 14 percent during the last visit (p-value = 0.28). The latter is relatively low, indicating that outright cheating was not very prevalent.[22] Given the extent of these behaviors, educators' scores from the school visits did not accurately reflect the quality of their teaching. Consequently, a decision was taken not to use the performance evaluation data to move forward or delay promotions.

We did not feel that it was appropriate to ask ASDEOs directly about this behavior in the qualitative interviews, and none of the respondents chose to speak openly about it. We can, however, explore this issue empirically. The second column in Table 4 reports results from a school-level regression of the proportion of videos in which coaching were detected) on school and classroom characteristics collected at baseline. There are no strong school or classroom predictors of cheating in test administration, with just one coefficient significant at the 10 percent level. By contrast, the second column in Table 5 reports several statistically significant associations between detection of cheating in test administration and ASDEO characteristics.[23] Detection is less likely at schools with a young ASDEO compared to schools with older ASDEOs, though the result is not very strong. It is also less likely at schools whose

[22] While we cannot quantify the extent of coaching of these behaviors in control schools as we lack videos, it is possible the improved teaching behaviors in control schools improved test scores in them.
[23] We aimed to capture personal relationships between educators and their ASDEO, asking if the ASDEO was a family member, neighbor, friend, or acquaintance. Only a small percentage reported a "close" relationship: 9% for head teachers and 4% for teachers. Due to limited variation, we excluded a "personal ties" variable from the regressions. The absence of such ties also suggests that a strategic motive for cheating—to help an educator get promoted —is unlikely.

ASDEO has lower public service orientation compared to schools with more motivated ASDEOs. We hypothesize that with a higher public service orientation, ASDEOs might feel a greater degree of responsibility to help teachers and students and be more likely to hint at answers. Some of the survey questions have to do with the trade off between doing something for the good of people and following the rules. Finally, detection is more likely at schools with an ASDEO with few schools to visit and more time to coach compared to schools with a busier ASDEO. Since fewer schools to visit could signal a more rural area or a less experienced inspector, however, it is hard to pin down a possible reason for this finding.

5.3 Spillovers into control schools

We believe that two aspects of the intervention may have spilled over from treatment to control schools, namely accountability pressure and improved understanding of teaching best practices. Given the design of the pilot program, we do not have videos of inspector behavior, so our evidence here comes from the qualitative interviews with E&SE officials.

All qualitative survey interviewees reported liking the new system and said it motivated educators to improve. In one interview, an E&SE official noted that teachers in control schools were asking for advice because they had been told by some ASDEOs that the pilot would eventually be rolled out in their schools as well. Even more tellingly, one female official noted, *"in the other [control] schools, we would take out our mobile and say this is not being recorded with pictures and stands, but we used to say that your voice is being recorded so that would make her afraid and give a good lesson and prepare well."* It appears that, excited by the treatments, ASDEOs attempted to replicate a form of accountability pressure in control schools, hoping this would improve educator behavior. For this reason, the RCT, as implemented, cannot speak to the efficacy of educator incentives.

Relatedly, we suspect that, having become familiar with the classroom observation tool in treatment schools, ASDEOs may have also adapted their coaching behaviors in control schools. Videos in treatment schools show that the aspects of pedagogy that ASDEOs were coaching on were precisely the "best practices" that scored highly in the rubric for the classroom observations tool—e.g., putting students into small groups. From the qualitative interviews, we know ASDEOs were committed to improving teaching in *all* schools. It therefore seems likely that ASDEOs did not stick to "business as usual" in control schools, but instead gave new guidance and feedback, much as we saw in treatment schools.

Together these two spillovers—accountability pressure and best practice coaching—provide a plausible explanation for why three of the six (pre-registered) educator outcomes improved in *all* study schools. Specifically, this supplementary evidence that ASDEOs took coaching into treatment schools and (aspects of) the new accountability regime into control schools may explain greater use of student groups, reporting progress, and head teacher help in all study schools.[24] Additionally, the teaching practices officials were coaching on were included in the classroom observations scoring rubric. There were no improvements in educator attendance, nor in helping students outside of school, however. We hypothesize that those outcomes may be higher-cost and lower-reward for educators. Splitting students into

[24] One might think that coaching occurring in both treatment and control schools means treatment and control schools differed only in whether an educator incentive program was in place—thus allowing us to isolate the effects of the incentive program. However, there are two problems with this. First, we cannot assume that the same amount of coaching occurred in treatment and control schools—and we indeed cannot say precisely how much more or less coaching occurred in one or the other. Per the design of the interventions, there are no videos in control schools, and from the qualitative interviews, we can simply say that coaching occurred in control schools as well. Second, because ASDEOs in control schools warned educators that the new system may arrive soon, without necessarily specifying when, control group educators were likely also influenced by the incentive program. We cannot quantify variation in the perceived strength of the incentives across educators in different treatment conditions but 30% of schools in the control group had heard about the program (compared to 96% in the treatment groups).

groups during class and keeping parents informed are potentially relatively easy for teachers to do; helping students outside of school may be harder. In many settings, it is not easy to regularly speak with parents. However, in our sample, villages are extremely small, and students almost exclusively attend the school closest to their home. Additionally, 58 percent of teachers live within 3 km of the school and 57 percent are originally from the village in which the school is located. Consequently, teachers can have natural and frequent encounters with parents. Teacher and head teacher attendance was already quite high, so further attendance improvements may have been perceived as too costly. Educators often have other duties (like attending trainings or meetings, helping with elections, etc.) and are allowed two excused absences per term. This norm thus also affected whether incentives would be successful (and had we been aware of it, we would not have incorporated this measure into the interventions).

One interpretation is that some ASDEOs found it difficult to change their 'mental model,' and continued to feel that it was their responsibility to provide advice. The 'business as usual' protocol suggests that a lot of interaction and coaching occurred. ASDEOs felt it was important to approach schools in a collaborative manner, with one official stating: *"when I go to a school, I go as a friend and representative and don't say that I'm an inspector and impose authority on them and [I] used to say that we are here to observe as a friend and if there is any discrepancy, we will correct it in a friendly manner."*

In the interviews, seven of eight officials expressed a desire to coach, with one noting: "*The most important role of the supervisor is to supervise the teaching, to see what the deficiencies in their teaching are and to point them out... [and] if they see something really good in one teacher, then they could convey that to the other teachers."* Another official noted that, *"We felt sometimes that we should say something to the teacher during the observation."*

The inspectors and some educators liked the pilot interventions; younger teachers in particular were pleased that they could be rewarded for their efforts and outcomes produced rather than only being promoted based on experience or connections. The E&SE Department heard such positive feedback about the pilot that partway through it, they began implementing a similar policy outside of our pilot administrative districts that focused on similar outcomes and used a similar process as ours, on tablets. However, the school visits under that program did not include rewards or sanctions. An official remarked that, *"in the [other] project that we are doing now, we have the option of giving feedback."* Note that we did not provide input into this new policy. Rather, officials appear to have "borrowed" some of our questions and measurement techniques. The above discussion provides a plausible explanation for the absence of treatment effects on teacher behavior or student outcomes.

5.4. Other Challenges

Alongside the challenges highlighted in Sections 4.1-4.3 that stem from the (largely well-intentioned) behavior of ASDEOs, the project also faced several political hurdles. First, there was high turnover with four different Education Secretaries. Additionally, the highly dynamic and supportive E&SE Department official initially tasked with overseeing the RCT was replaced. Such turnover is not abnormal but did drastically affect the integrity of implementation. Second, ASDEOs proved unable to do as many school visits as planned because they were overwhelmed with other responsibilities and were not well resourced. Third, an initial proposal to provide incentives for ASDEOs, to complement those for educators, was watered down. Rather than offering explicit career incentives, the provincial government ultimately agreed simply to put a note in ASDEOs' promotion dossiers regarding the quality of their visits. Each of these challenges reduced the quality and fidelity of the implementation of the pilot and likely weakened the effects of the program.

6 Conclusion

This paper presents insights into the challenges of collaborating with government to co-design and integrate interventions into local bureaucratic systems. We offer four pieces of advice for researchers considering such collaboration, gleaned from an RCT in Pakistan.

First, it is advisable to secure a detailed MOU at the outset describing the design *and implementation* of the program. While we had an MOU with the Education Secretary when we started the project, four different secretaries were appointed during the project and the MOU was quite general. It did not provide many project or implementation details, such as who had what responsibilities, how intervention fidelity would be tracked, or how communication would occur during the study. The research team thus had little recourse and not always the right points of contact. Specifying this in greater detail may prove challenging (e.g., if government is unable to commit to design parameters before costs or political feasibility are fully known) but may help ensure a shared understanding of project needs.

Second, consider investing in qualitative work and piloting specifically geared toward understanding likely implementation challenges *in advance*. Speaking with stakeholders in non-study areas and a very small-scale pilot can help substantially. Insights may surface where interventions cut against prevailing norms and practices—such as the request for ASDEOs to forsake coaching during monitoring visits—and how, if at all, the government or the research team might circumvent these concerns to ensure effective implementation.

Third, while it can be beneficial to leverage existing systems and staff, be cautious if asking a civil servant to simultaneously take on distinct roles. Officials may deviate from new protocols (despite training) if continually immersed in familiar ways of doing things. In our study, the roles of active coach and passive performance monitor appeared to entail

distinct "mental models" of one's job, and ASDEOs were uncertain how to switch effectively between models for control and treatment schools. It may be necessary (and maybe a red line for program design) for different officials to take on distinct roles if a new role is to be tested.

Fourth, consider whether salient incentives are necessary to encourage adherence to new protocols, and whether training needs to be continuous (e.g., with reminders and check-ins), or more extensive. For example, in addition to detailed training on *what* must be done, it may be useful to engage civil servants in discussions of *how* they will change their mental model (e.g., overcome the tendency to fall into old habits).[25] Researchers may also want to monitor implementation to confirm the right protocols are used—e.g., through spot checks, reporting requirements, or other built-in safeguards and feedback mechanisms.

The implementation challenges that bedeviled the program we describe resulted in a missed opportunity to learn about the efficacy of educator incentives. Yet all was not lost; thanks to insights from video recordings and qualitative interviews, we have an unusually detailed account of *why* the interventions failed. These lessons will hopefully inform future studies seeking to co-design with, and embed interventions in, local bureaucratic systems.

[25] Old habits, while not inherently negative, may, in the context of an RCT, blur the distinction between treatment and control groups. For evidence on the effectiveness of coaching, see Cilliers et al. (2020).

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. 2017. "When Should You Adjust Standard Errors for Clustering?" NBER Working Paper (No. w24003).

Anderson, Michael. L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." Journal of the American Statistical Association 103(484): 1481-1495.

Araujo, María Caridad; Rubio-Codina, Marta; Schady, Norbert Rüdiger. 2021. 70 to 700 to 70,000: Lessons from the Jamaica Experiment. The Scale-Up Effect in Early Childhood & Public Policy: Why Interventions Lose Impact at Scale and What We Can Do About It. Routledge.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." Journal of Economic Perspectives 31(4): 73-102.

Barlevy, Gadi, and Derek Neal. 2012. "Pay for Percentile." American Economic Review 102(5): 1805-1831.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, and Justin Sandefur. 2018. "Experimental Evidence on Scaling Up Education Reforms in Kenya." Journal of Public Economics 168: 1-20.

Carter, Samantha, Iqbal Dhaliwal, Sam Friedlander, and Claire Walsh. 2021. Forging Collaborations for Scale: Catalyzing Partnerships among Policy Makers, Practitioners, Researchers, Funders, and Evidence-to-Policy Organizations. In The Scale-Up Effect in Early Childhood and Public Policy, pp. 370-388. Routledge.

Cilliers, Jacobus, Ibrahim Kasirye, Clare Leaver, Pieter Serneels, and Andrew Zeitlin. 2018. “Pay for Locally Monitored Performance? A Welfare Analysis for Teacher Attendance in Ugandan Primary Schools.” Journal of Public Economics 167: 69-90.

Cilliers, Jacobus, Brahm Fleisch, Christel Prinsloo, and Stephen Taylor. 2020. “How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching.” Journal of Human Resources 55(3): 926-962.

Ganimian, Alejandro J. 2020. Growth-Mindset Interventions at Scale: Experimental Evidence from Argentina. Educational Evaluation and Policy Analysis 42(3): 417-438.

Gilligan, Daniel O., Naureen Karachiwalla, Ibrahim Kasirye, Adrienne M. Lucas, and Derek Neal. 2019. “Educator Incentives and Educational Triage in Rural Primary Schools.” Journal of Human Resources 57(1): 79-111.

Habib, Masooma. 2015. “Teacher and School Administrator Incentives for Improved Education Delivery in Khyber Pakhtunkhwa Province, Pakistan.” IGC Working Paper.

Holmstrom, Bengt, and Paul Milgrom. 1991. “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.” Journal of Law, Economics, and Organization 7: 24-52.

Karachiwalla, Naureen, and Albert Park. 2017. “Promotion Incentives in the Public Sector: Evidence from Chinese Schools.” Journal of Public Economics 146: 109-128.

List, John A., Dana Suskind, and Lauren H. Supplee (Eds.). 2021. The Scale-Up Effect in Early Childhood and Public Policy: Why Interventions Lose Impact at Scale and What We Can Do About It. Routledge.

List, John. A. 2022. The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale. Penguin Random House.

McKenzie, David. 2012. “Beyond Baseline and Follow-Up: The Case for More T in Experiments.” Journal of Development Economics 99(2): 210-221.

Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal. 2017. “The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India.” Journal of Public Economics 145: 116-135.

Muralidharan, Karthik, and Abhijeet Singh. 2023. “Improving Public Sector Management at Scale? Experimental Evidence on School Governance in India.” Working Paper.

Pakistan Public Administration Research Centre. 2004. A Guide to Performance Evaluation. Islamabad.

Pappas, Sophia E. 2021. “Real-World Application and Understanding of the Threats to Scaling: Commentary 1: Chapters 7, 8, 9, and 10.” In The Scale-Up Effect in Early Childhood and Public Policy, pp. 199-210. Routledge.

Schaffner, Julie, Paul Glewwe, and Uttam Sharma. 2021. “Why Programs Fail: Lessons for Improving Public Service Quality from a Mixed-Methods Evaluation of an Unsuccessful Teacher Training Program in Nepal.” Working Paper.

Steinberg, Matthew P., and Morgaen L. Donaldson. 2016. “The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era.” Education Finance and Policy 11(3): 340-359.

Tanwir, Maryam. 2010. Bureaucratic Perception of Merit, Gender and Politics. University of Cambridge.

Tanwir, Maryam, and Azam Chaudhry. 2015. “The Performance Evaluation System in Pakistan’s Civil Service.” Africa’s Public Service Delivery & Performance Review 3(2): 81-103.

World Bank. 2015. Conducting Classroom Observations: Analyzing Classroom Dynamics and Instructional Time, Using the Stallings' Classroom Snapshot Observation System. User Guide.

Table 1: Intervention Implementation

Variable	Control	Teacher treatment	Headteacher treatment
Experienced "more" visits than usual		11%	8%
Experienced "far more" visits than usual		81%	82%
Heard about the program	30%	93%	95%
Number of visits by ASDEO this year	3-6	2.7	2.6
Duration of visits	50 minutes	7 hours	7 hours
The new visits would measure:			
Own attendance		66%	60%
Teacher pedagogy		70%	54%
Student attendance		76%	61%
Student learning (incorrect in the head teacher treatment)		56%	47%
Good performance would increase chance of promotion		79%	67%
Salary would increase without promotion (incorrect)		23%	25%

Notes: Data about the implementation of the interventions was not collected from control schools at endline. The number and duration of visits reported in the control group come from baseline qualitative information collected from inspectors. For the teacher and headteacher treatments, the data were collected in the endline survey individually at each school.

Table 2: Impact estimates of teacher and head teacher interventions

Panel A: Student Outcomes			
	Proportion Correct	Student Present	Dropped Out
	(1)	(2)	(3)
Head Teacher Treatment	-0.004	-0.030	0.002
	(0.008)	(0.016)	(0.004)
Teacher Treatment	0.006	-0.029	-0.001
	(0.009)	(0.016)	(0.002)
R^2	0.349	0.089	0.032
Mean control group	0.485	0.857	0.003
Observations	4,639	6,389	6,407
Panel B: Teacher Pedagogy			
	Teacher Helps Individually Outside of School	Teacher Splits Students into Groups	Teacher has Spoken to Parents about Progress
	(1)	(2)	(3)
Head Teacher Treatment	-0.009	-0.008	-0.039
	(0.041)	(0.043)	(0.041)
Teacher Treatment	0.044	0.023	0.040
	(0.044)	(0.043)	(0.043)
R^2	0.235	0.157	0.062
Mean control group	0.579	0.673	0.389
Observations	4,644	4,644	4,644
Panel C: Educator Effort			
	Teacher Present	Head teacher Present	Head teacher Helps G4 Math
	(1)	(2)	(3)
Head Teacher Treatment	-0.001	-0.035	-0.061
	(0.035)	(0.024)	(0.055)
Teacher Treatment	0.025	-0.061*	-0.064
	(0.031)	(0.026)	(0.055)
R^2	0.065	0.120	0.114
Mean control group	0.950	1.000	0.887
Observations	239	239	239

*Notes: * p-value<0.1, ** p-value<0.05, *** p-value<0.001. Standard errors in parentheses. Results derived from a regression of the outcome on the two treatment variables, the baseline value of the outcome variable, strata dummy variables, and standard errors are clustered at the level of the school and are unadjusted. All p-values used to determine statistical significance are adjusted for multiple hypothesis testing using Anderson's family-wise error rate adjusted q-values within panels. Panel A: proportion of questions answered correctly by grade 4 students on endline written grade 4 math assessment (of 45 questions); indictor for grade 4 student present at endline survey (measured by the enumerator); indicator for grade 4 student dropped out of school between the baseline to endline surveys (measured by the enumerator). The baseline control variable here is an indicator for a student who enrolled in school but dropped out before the baseline survey. Panel B: indicator for grade 4 student reporting that their teacher helps them individually outside of school, splits students into groups in class, and speaks to their parents regarding their progress in math. Panel C: indicator for teacher present on the day of the endline survey; indicator for head teacher present on the day of the endline survey; indicator for head teacher helping with the grade 4 math class.*

Table 3: Incidence of inspector behaviors observed in video recordings

Type of inspector behavior	Overall	Teacher treatment (T)	Headteacher treatment (HT)	p-value (T vs HT)	Visit 1 (V1)	Visit 3 (V3)	p-value (V1 vs V3)
Panel A: Classroom Observation							
Prompting teacher specific activity	18%	14%	23%	0.15	13%	9%	0.32
Telling teacher to use English	1.25%	0%	2.5%	0.16	1%	1%	-
Any of the above	19%	14%	24%	0.11	11%	7%	0.24
Panel B: Oral math test							
Gave student answer	8%	8%	-	-	4%	14%	0.28
Hinted the answer to student	36%	36%	-	-	21%	26%	0.37
Someone was present in the class	20%	20%	-	-	11%	11%	-
Any of the above	36%	36%	-	-	36%	26%	0.01
Number of schools	160	80	80		80	80	

Note: No oral math tests were given in head teacher treatment schools. The percentages in the columns representing all schools, T schools, and HT schools are based on indicator variables equal to one if the particular behaviour was uncovered in a school during either the first or second visit. The percentages in the V1 and V3 columns are based on indicator variables equal to one if the particular behaviour was uncovered in a school in that specific visit.

Table 4: Predictors of coaching and cheating – baseline school characteristics

	(1) Class Observation	(2) Oral math test
Head Teacher treatment	0.049	
	(0.037)	
Girls' school	-0.013	0.693
	(0.085)	(0.432)
Head Teacher Experience	0.002	0.001
	(0.002)	(0.005)
Head Teacher has B.A.+	-0.005	0.101
	(0.048)	(0.122)
Head Teacher Public Service orientation	0.037	-0.038
	(0.032)	(0.051)
Student-teacher ratio	0.002	0.008*
	(0.002)	(0.005)
Total enrollment	-0.000	-0.000
	(0.000)	(0.001)
Index of school resources	-0.004	-0.003
	(0.016)	(0.046)
Student score on written math exam	-0.003	0.013
	(0.004)	(0.008)
Student asset index	0.045	0.093
	(0.032)	(0.113)
Distance (in KM) of school from DEO office	-0.003	-0.023
	(0.001)	(0.019)
R^2	0.461	0.579
Number of schools	142	67

Notes: The outcome variable in column 1 is the proportion of videos in a school in which coaching on the pedagogy assessments was detected. The outcome variable in column 2 is the proportion of videos in a school in which coaching on the math assessments was detected. Regressions include administrative district fixed effects and standard errors are clustered at the school level. Note that oral math tests were only administered in headteacher treatment schools.

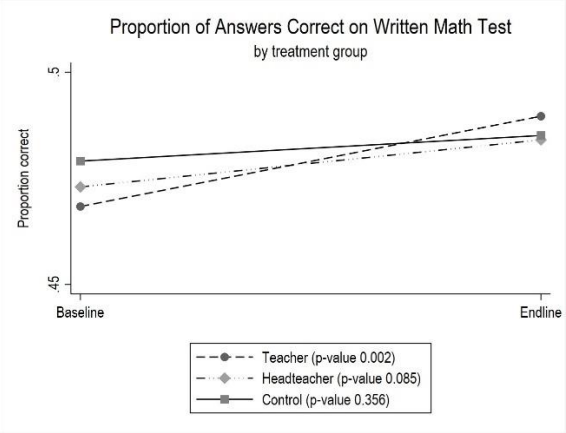
Table 5: Predictors of coaching and cheating – baseline inspector characteristics

	(1) Class Observation	(2) Oral math test
Head Teacher treatment	0.092	
	(0.059)	
Inspector's age is below median	0.056	-0.216*
	(0.063)	(0.123)
Inspector's experience below median	0.152	0.252
	(0.147)	(0.248)
Inspector is female	0.265**	-0.078
	(0.106)	(0.160)
Inspector's family owns land	0.097	-0.033
	(0.085)	(0.164)
Inspector's mother is from same Tehsil as school	-0.105	0.119
	(0.080)	(0.162)
Number of schools inspector is responsible for below median	-0.005	0.056**
	(0.062)	(0.121)
Inspector's public service orientation	0.038	-0.309***
	(0.085)	(0.153)
R^2	0.206	0.119
Number of schools	152	76

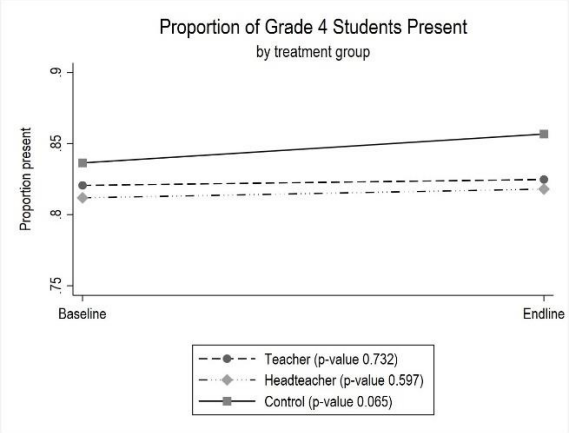
Notes: The outcome variable in column 1 is an indicator variable equal to one if there was any coaching detected in a school in any video that was watched. The outcome variable in column 2 is an indicator variable equal to one if a district official either hinted at or gave a student the answer to a question on the oral math test during any video. Regressions include administrative district fixed effects and standard errors are clustered at the school level. Note that oral math tests were only administered in headteacher treatment schools. Demographic information is missing for one district official, rendering the number of observations smaller than the total number of schools.

Figure 1: Student Outcomes

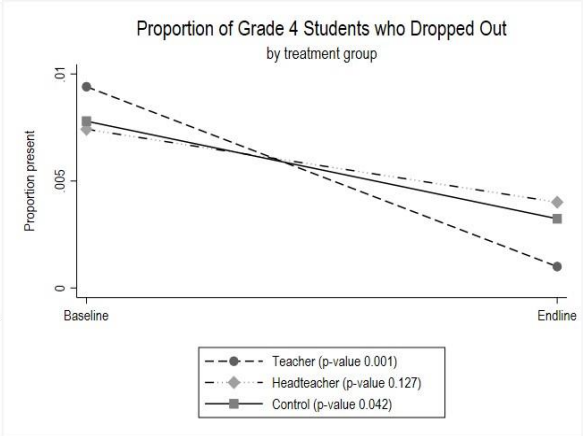
Panel A



Panel B



Panel C

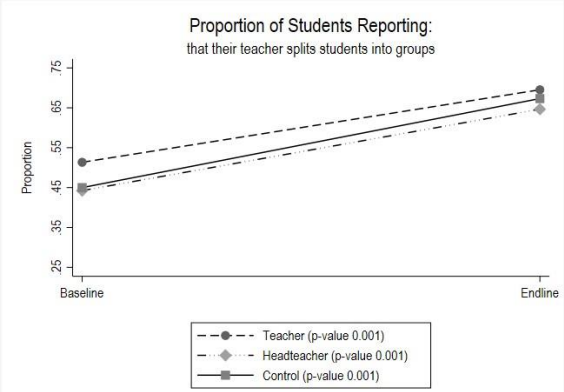


Notes: Paired t-tests are used to calculate the p-value of the difference between the baseline and endline values of the outcomes, separately across the three groups. These are reported in parentheses next to the group labels. In Panel C, the baseline value of dropout is calculated as the proportion of students who were enrolled in school at the beginning of the school year but dropped out of school by the time the baseline survey occurred. The endline value of dropout is calculated as the proportion of students who dropped out of school between the baseline and endline surveys.

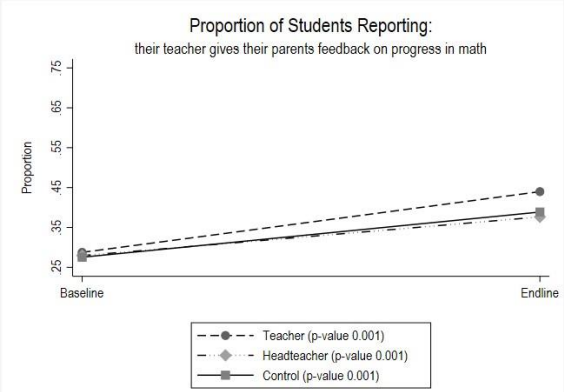
Figure 2: Educator Outcomes

Teacher Practices

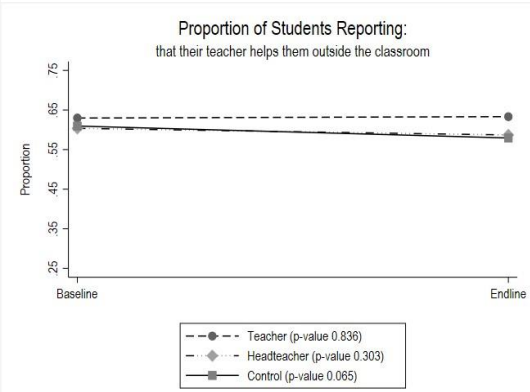
Panel A



Panel B

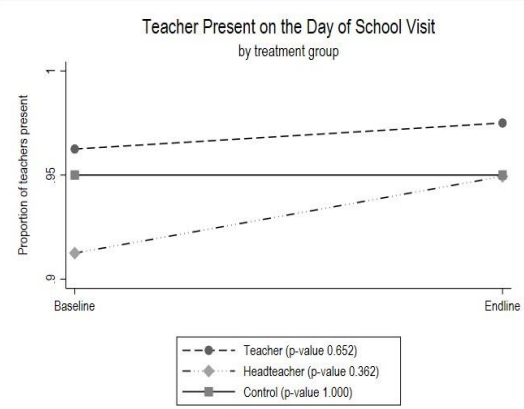


Panel C

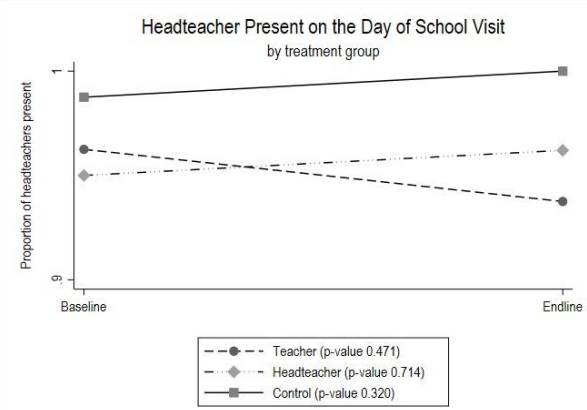


Educator Effort

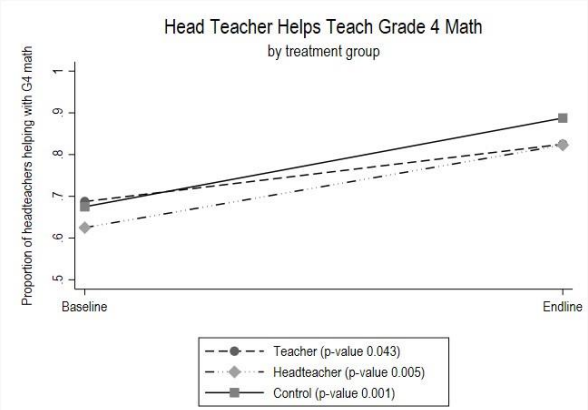
Panel D



Panel E



Panel F



Notes: Paired t-tests are used to calculate the p-value of the difference between the baseline and endline values of the outcomes, separately across the three groups. These are reported in parentheses next to the group labels.