INTERNATIONAL
FOOD POLICY
RESEARCH
INSTITUTE

PAKISTAN

RESEARCH
PROGRAM ON
Policies,
Institutions,
and Markets

Led by IFPRI

IFPRI Discussion Paper 01987

December 2020

Coaching, not monitoring?

Insights from a school accountability reform in Pakistan

Naureen Karachiwalla

Katrina Kosec

Saher Asad

Masooma Habib

Clare Leaver

Attique ur Rehman

Development Strategy and Governance Division

Poverty, Health, and Nutrition Division

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

AUTHORS

Naureen Karachiwalla* (n.karachiwalla@cgiar.org) is a Research Fellow in the Poverty, Health, and Nutrition Division of the International Food Policy Research Institute (IFPRI), Washington DC.

Katrina Kosec (k.kosec@cgiar.org) is a Senior Research Fellow in the Development Strategy and Governance Division of the International Food Policy Research Institute (IFPRI), Washington DC.

Saher Asad (saher.asad@lums.edu.pk) is an Assistant Professor in the Economics Department at the Lahore University of Management Sciences (LUMS), Lahore, Pakistan.

Masooma Habib (mhabib@gwmail.gwu.edu) is a Research Fellow at the Consortium for Development Policy Research (CDPR), Lahore, Pakistan.

Clare Leaver (clare.leaver@bsg.ox.ac.uk) is an Associate Professor of Economics and Public Policy at the Blavatnik School of Government and Fellow of University College Oxford.

Attique Ur Rehman (aatique.rehman@gmail.com) is a PhD candidate at the University of Delaware and an Adjunct Faculty member at the Lahore University of Management Sciences (LUMS), Lahore, Pakistan.

** Corresponding author*

Notices

¹ IFPRI Discussion Papers contain preliminary material and research results and are circulated in order to stimulate discussion and critical comment. They have not been subject to a formal external review via IFPRI's Publications Review Committee. Any opinions stated herein are those of the author(s) and are not necessarily representative of or endorsed by IFPRI.

² The boundaries and names shown and the designations used on the map(s) herein do not imply official endorsement or acceptance by the International Food Policy Research Institute (IFPRI) or its partners and contributors.

³ Copyright remains with the authors. The authors are free to proceed, without further IFPRI permission, to publish this paper, or any revised version of it, in outlets such as journals, books, and other publications.

Coaching, not monitoring? Insights from a school accountability reform in Pakistan*

Naureen Karachiwalla, Katrina Kosec, Saher Asad, Masooma Habib, Clare Leaver,
and Attique ur Rehman¹

December 21, 2020

Abstract

We evaluate a pilot reform of the promotion system for teachers and head teachers in Khyber Pakhtunkhwa province, Pakistan, that was implemented as a randomized controlled trial during the 2017-2018 school year. District officials were asked to visit treatment schools to monitor educator behavior and student performance passively. Educators in treatment schools were told that their relative performance on these measures would determine their promotion speed. District officials were asked to continue with 'business as usual' visits in control schools, during which they often coached educators. We find null treatment effects for all pre-registered student and educator outcomes. In-class teaching practices improved from baseline to endline, but these sizeable changes were similar in magnitude in *all* schools. Videos of inspection visits and follow-up interviews reveal that district officials coached teachers on these practices in all schools, indicating that they struggled to adapt their 'mental model' to the new accountability regime.

JEL Codes: H1, H83, I25, I28

* 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 Kaiani, as well as excellent survey management from RCONS. We also thank Jeffrey Patton for invaluable advice regarding psychometrics. Funding for this study was provided by the CGIAR Research Program on Policies, Institutions, and Markets led by IFPRI; 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.

¹ Karachiwalla (corresponding author) and Kosec: International Food Policy Research Institute. Asad: Lahore University of Management Sciences. Habib and Rehman: Consortium for Development Policy Research. Leaver: Blavatnik School of Government, University of Oxford.

1 Introduction

Promotion systems can, in principle, generate implicit incentives: an employee is motivated to work hard to impress an evaluator today because she cares about a promotion tomorrow. Yet, in practice, such systems do not work well in many government bureaucracies. Career incentives are often weak because evaluators lack informative performance measures, or simply because promotions are influenced by seniority, political connections, or other factors unassociated with performance (Bana and McCourt, 2006).

This paper evaluates a pilot reform of the promotion system for teachers and head teachers that was instigated by the Elementary and Secondary Education (E&SE) Department in Khyber Pakhtunkhwa (KP) province, Pakistan (Habib, 2015). The reform sought to ensure that the E&SE Department collected objective and informative performance measures that could be used to determine teacher and head teacher promotions. The aim was to improve educator effort and teaching practices and, ultimately, student outcomes.

The pilot implemented two different interventions during the 2017-2018 school year in rural public schools in the districts of Charsadda, Mardan, and Nowshera. In the first intervention, focused on grade 4 math *teachers*, an official from the E&SE Department was asked to make three unannounced visits to schools to monitor the teacher's attendance, the attendance of the teacher's students, the teacher's pedagogy, and the learning levels of this teacher's students. The official was given a tablet and asked to take photos to document attendance and record videos to document classroom observations (of teacher pedagogy) and orally administered learning assessments. The second intervention, focused on *head teachers*, was identical except that the E&SE official was asked to monitor the head teacher's attendance, attendance of all teachers in the school, attendance of all students in the school, and the pedagogy of two teachers. At the start of the school year, teachers and head teachers in the treatment groups were informed that their scores during inspection visits would be compared to those of teachers and head teachers in the same school district, and relative performance would determine the speed of their promotions. Treatment assignment occurred at the school level. In each of 20 school districts, four schools were randomly assigned to each of the teacher treatment, the head teacher treatment, and the control group (in which 'business as usual' school visits occurred), yielding a sample of 240 schools. Importantly for what follows, in each school district, the same E&SE official conducted the inspection visits in the eight

treatment schools *alongside* the 'business as usual' visits, produced prior to the intervention) in the four control schools.

We report results for all student and educator outcomes in our trial pre-registration. We find no evidence that the interventions improved student learning, attendance, or dropout relative to the control at the student-level. These outcomes are balanced at baseline, and there is no economically meaningful improvement in any study arm five months later, at endline.

At the educator-level, we find no evidence that the interventions improved in-class practices (the teacher splitting students into groups in the classroom, the teacher informing parents about their child's math progress, or the head teacher helping teach the grade 4 math class) relative to the control. However, for each of these outcomes, there is a statistically significant improvement from baseline to endline that is of similar magnitude in all three study arms. Effect sizes are non-trivial: we observe a 15-20 percentage point increase in the likelihood of a student's 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 with teaching the grade 4 math class, for all three study arms.

We find no evidence that the interventions improved out-of-class behavior (teacher attendance, head teacher attendance, or teachers helping students outside of class) relative to the control. Teacher attendance and help outside of class are balanced at baseline, and there is no meaningful improvement in any study arm by endline. Head teacher attendance is balanced at baseline and improves slightly in the control arm while declining slightly in the teacher arm by endline, producing a weakly significant and negative treatment effect.

To explore possible mechanisms for the in-class practice improvements, we watched the classroom observation videos of teachers recorded during classroom observations in treatment schools. These videos show that, rather than switching to the role of passive performance monitor as the pilot intended, many E&SE officials continued to use the same active coaching practices during their visits as they had prior to the intervention. In a quarter of treatment schools, there is documentary evidence of officials coaching teachers to split students into groups and use more learning materials. To gain further insights into this behavior, we conducted a series of semi-structured interviews with a sub-sample of 6 of the 20 E&SE officials. During these interviews, and without prompting, officials gave examples of their coaching of teachers, both during and after pedagogy observations, and explained how they wanted to continue to give teachers feedback in treatment schools. One official also spoke

openly about adopting aspects of the interventions in control schools. In short, E&SE officials continued to coach in treatment schools (as they were doing in control schools) and tried to take elements of the accountability system from treatment schools into control schools. This behavior provides a plausible explanation for our finding of similar in-class practice improvements in all study schools (hence the absence of treatment effects). Of course, the question remains as to why these improvements in teaching practices only translate into small learning gains in all schools. We suspect that the short duration of the pilot is part of the answer.

In many trials with null results, researchers must conjecture underlying mechanisms. Here, thanks to the videos, it is unambiguous that the intervention was not implemented as intended. This evidence makes it easier to draw policy recommendations. The E&SE Department decided from the outset that the officials who are currently tasked with visiting schools (to provide coaching and feedback) should undertake the new inspections. This choice was made to ensure that the pilot interventions were embedded within the system. While a laudable aim, our results highlight a potential downside of this decision. The roles of an *active coach* and a *passive performance monitor* are not the same and likely entail very different 'mental models' of what one's job entails. It appears that some officials either found it difficult or did not want, to switch between them.

The natural policy question is whether the government should work towards a more coherent accountability system with strengthened training and/or credible incentives for performance monitors or harness the apparent enthusiasm for coaching by supporting government officials to provide more systematic coaching and mentoring at the frontline. In fact, these policies are not mutually exclusive, and the government could consider both. The key takeaway from this paper is that the roles of an active coach and passive performance monitor are strategically different and may need to be performed by separate personnel.

To our knowledge, this is the first paper to study the impact of promotion incentives on the behavior of frontline bureaucrats employing a randomized controlled trial. In contrast to the null treatment effects reported here, Karachiwalla and Park (2017) use observational data and provide evidence that teachers in China responded to the prospect of promotion.¹ There is also evidence of the effects of promotion systems for academics in high-income settings (Haeck and Verboven, 2012; Nieddu and Pandolfi, 2018).

¹ In China, all civil servant promotions are based on extensive performance evaluations. This system has been in place for decades, so civil servants and evaluators have been acting as passive monitors for a long time.

More broadly, our paper is related to the literature on the design of accountability systems in education and, in particular, the unintended effects that such systems can have depending on their design.² For example, Figlio and Winicki (2005) find that educators in the United States manipulate test-based accountability systems by providing high-calorie meals on test days and strategically encouraging weak students to stay home.³ Both Gilligan et al. (2019) and Cilliers et al. (2020) note that using high-powered incentives for schools to improve their national exam scores encourages triage behavior whereby weak students are encouraged to drop out of school before taking the exam. Glewwe, Ilias, and Kremer (2010) find that teachers in Kenya "teach to the test" (on which their performance will be assessed) instead of creating lasting learning gains for students. Our study shows how the impact of an accountability system based on a range of performance measures (beyond student test scores) also hinges on design. Rather than the 'gaming' in the above papers, our results point to a different phenomenon—the stickiness of mental models—that arises as existing officials are given new accountability roles.

The remainder of the paper is organized as follows: Section 2 describes the interventions we evaluate. Section 3 presents our empirical specification, and Section 4 characterizes our results. Section 5 provides some discussion. Finally, Section 6 concludes.

2 Intervention

Prior to the pilot, both teachers and head teachers had incentive systems in place that should tie performance to promotions *de jure* (see Appendix Section A.1 for further details). There were annual evaluations for teachers, but these were not tailored to teaching, covered the calendar year rather than the school year, and were conducted by the head teacher, for whom impartiality was reportedly challenging. Head teachers were evaluated via school inspection visits, but these were unstructured and did not monitor aspects relevant to learning (including learning itself). Importantly, neither the teacher evaluations nor the inspection visits were tied to careers. Promotions were instead based on seniority and connections, providing no *de facto* incentives.

² See Neal (2011) for an overview of the evidence in high-income countries and Glewwe and Muralidharan (2016) for evidence from low-income countries.

³ See also Figlio and Getzler (2006).

The pilot sought to strengthen teacher and head teacher career incentives by reforming evaluation and promotion processes. The two interventions were designed in close collaboration with the KP E&SE Department and built on theory outlining how to incentivize educators effectively (Barlevy and Neal, 2012). They were implemented as a randomized controlled trial (RCT) in 240 public, rural primary schools in the districts of Charsadda, Mardan, and Nowshera during the 2017-2018 school year (see further intervention details in Appendix Section A.2, timeline in Appendix Figure A1, and sampling details in Appendix Section A.3).⁴ Baseline and endline surveys conducted at the beginning and end of the school year gathered data on grade 4 math teachers, grade 4 students, head teachers, and schools. For tractability and to facilitate impact evaluation, the pilot targeted grade 4 students and their math teacher and the school's head teacher. Appendix Section A.2 also describes how this choice of cohort and subject permitted the collection of reliable and comparable data while retaining policy relevance.

In each of 20 school districts, 12 schools were randomly assigned to one of three equally sized study arms: a reformed teacher evaluation in which an E&SE official implemented the reformed annual evaluation (the teacher treatment); a reformed head teacher inspection in which an E&SE official conducted the reformed school inspection visits (the head teacher treatment); and a control group in which an E&SE official conducted 'business as usual' school visits. Details on treatment assignment and power calculations appear in Appendix Section A.4.

Schools assigned to the *teacher treatment* (T) were scheduled to receive three unannounced inspection visits (assigned by the research team) collecting data on teaching-specific outcomes—the presence of the teacher and of his/her students, the teacher's pedagogy (measured using a Stallings (1977) type classroom observation tool), and the test scores of the teacher's students on an oral math test (test construction details contained in Appendix Section A.5). Scores overall visits would be aggregated into a single, cardinal score for each teacher and explicitly linked to career progression via promotion tournaments within school districts. Teachers who performed the best relative to others in the same school district and treatment arm would have their promotion fast-tracked by a year, while those who performed worst would have their promotion delayed by a year (the middle teachers were unaffected).

⁴ The school year in this context is short at 5 months (not including 2 months of school enrolment drives in May and September). The intervention spanned 4 months, so it could end before exams began.

Schools assigned to the *head teacher treatment* (HT) were also scheduled to receive three unannounced inspection visits. The information collected was the presence of head teacher, all teaching staff, and all students within the school, and the pedagogy of two teachers via classroom observations. The same promotion tournaments affected the head teacher's promotion. As the same number of educators' promotions would be fast-tracked as delayed, these interventions were budget neutral. Detailed scoring rubrics are described in Appendix Section A.2 and were distributed to grade 4 teachers and head teachers during the first surprise inspection visit. Educators in the control group were told nothing. Each of the four metrics within each intervention were recorded on tablets using photos (for attendance) and videos (for classroom observations of teacher pedagogy and administering the oral math test to students). The tablets required a project-designated school ID code to access the system, preventing their use in control schools.⁵

Crucially, the same E&SE official was responsible for visiting all study schools in a circle.⁶ Prior to the roll-out of the interventions, these officials were asked to prepare a 'business as usual' protocol describing their current practices during school visits. These protocols generally point to active engagement and coaching with educators and children. E&SE officials received training on the distinct, new monitoring role they were to play in treatment schools and were instructed to continue 'business as usual' in control schools.

3 Empirical Strategy

We estimate reduced form equations to uncover intent to treat impact estimates:

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

where Y_{it} is an outcome for unit i , which will either be a student or a school, at time t (the baseline was in October 2017, and the endline was in February 2018). T_i and HT_i are indicator variables for schools assigned to the teacher treatment and head teacher treatment, respectively. We include $Y_{i,t-1}$ as a control for the baseline value of the outcome, given the low

⁵ The E&SE Department wanted to involve a range of officials in inspections: District Education Officers and their deputies who work at the district level (one administrative unit below the province), Sub-District Education Officers who work at the Tehsil level (the administrative unit below the district), Assistant Sub-District Education Officers at the circle level (equivalent to U.S. school districts). In practice, ASDEOs performed the vast majority of inspections.

⁶ Each school district is gender-segregated so there is a female inspector and a male inspector for schools of different genders in the same area. Inspectors only visited schools from their own district.

level of autocorrelation, these outcomes tend to have (McKenzie, 2012). Circle fixed effects, γ_c , are included as randomization was stratified by circle. Standard errors are clustered at the school level as treatment was assigned by school (Abadie et al., 2017).

We group outcomes into families of related outcomes (student outcomes, teacher practices, and educator effort) and use multiple hypothesis testing to correct for the number of tests conducted within a family to ensure that our estimates are not statistically significant merely by chance. We use the family-wise error rate described by Anderson (2008) and report statistical significance using these adjusted p -values. Our coefficients of interest are β_1 and β_2 , where β_1 is the impact of the teacher treatment and β_2 is the impact of the head teacher treatment.

4 Results

4.1 Sample characteristics and balance

Appendix Table A1 presents student-level summary statistics (Columns 1, 3, and 5) and p -values from tests for balance (Columns 7-9) between the teacher treatment, head teacher treatment, and control groups. Appendix Table A2 shows summary statistics and p -values from tests for balance for school-level outcomes and follows the same format as Appendix Table A1. In each table, only one test in 39 reveals a statistically significant imbalance.

4.2 Attrition

To check for selective attrition among students, teachers, and head teachers, we regress an indicator variable equal to one if the student, teacher, or head teacher dropped out of the study between baseline and endline.⁷ Appendix Table A3 presents the results, which are at the school level. There is no differential attrition by treatment group for students (Column 1) or teachers (Column 2), but there is a difference for head teachers (Column 3). Head teachers in the teacher treatment arm are more likely to attrit.

⁷ One school merged with a non-sample school by the time of the endline survey, so we did not interview that school at endline, and our final sample comprises 239 schools for the analysis.

4.3 Student-level outcomes

We examine three student-level outcomes: the proportion of questions answered correctly on an endline written grade 4 math assessment (out of 45), whether the student was present on the day of the endline survey, and an indicator variable for dropping out of school between baseline and endline. These outcomes were pre-specified.⁸

It is instructive to begin by looking at the evolution of these outcomes over time. Figure 1 presents the means of each of these variables, at baseline and endline, and the p -values from paired t -tests for differences between baseline and endline means for each of the three study arms. Panel A reveals statistically significant improvements in test scores in both treatment groups and a similarly sized, but statistically insignificant, improvement in the control group.⁹ However, the magnitude of these changes is trivial—about one more question correct. Panels B and C tell a similar story. There is a statistically significant improvement in student attendance in the control group, but no economically meaningful change in any study arm. There are statistically significant declines in dropout in the teacher treatment and control groups, but no economically meaningful change in any study arm.

We estimate Equation (1) for the same student outcomes, reporting results in Table 1, Panel A. We present the coefficient, standard error, and level of significance for each of our two treatments using a p -value corrected for multiple hypothesis testing.¹⁰ As one would expect given Figure 1, we find no significant treatment effects on any student outcomes; coefficients reflect precisely estimated zeros.¹¹

4.4 Educator-level outcomes

We consider six educator outcomes: three related to teacher practices, and three pertaining to educator effort, also all pre-registered. For teacher practices, we test whether grade 4 students differentially report that their teacher: splits students into groups during class; speaks to their

⁸ <https://www.socialscisearch.org/trials/2815>.

⁹ Student test scores using norm-referenced scores derived from Item Response Theory (Hambleton, Swaminathan, and Rogers, 1991) produce the same results.

¹⁰ We use Anderson’s family-wise error rate adjusted q -values within panels of Table 1 (Anderson, 2008).

¹¹ The impact of the head teacher treatment on endline test scores is -0.004 percentage points and is not statistically significant. The 95 percent confidence interval [-0.011, 0.023] includes zero. Given the confidence interval, we can rule out with 95 percent confidence, an effect size larger than 0.023, or a difference of 1.2 percentage points. This effect size is also less than the effect we were powered to detect of 0.1. Thus, we interpret this effect as a precisely estimated null. For the teacher treatment, the impact estimate is 0.006 percentage points and is also not statistically significant. Once again, the 95 percent confidence interval [-0.021, 0.0129] includes zero, we are able to rule out effect sizes larger than 0.013 with 95 percent confidence (again, incredibly small), and the largest effect size is below the effect size we were powered to detect. Thus, we consider this impact estimate a precisely estimated zero as well.

parents regarding their progress in math; and helps them with classwork individually outside of school. These are reported at the student level and represent the proportion of students reporting this teacher behavior. For educator effort, we examine indicators for the grade 4 math teacher being present on the day of the endline survey; for the head teacher being present that day; and for the head teacher helping with teaching the grade 4 math class at endline. These are reported at the teacher or head teacher level (239 observations for 239 schools).

Figure 2 displays outcome means at baseline and endline for each study arm. There are statistically significant and economically meaningful improvements over time for three of the six outcomes: use of student groups in class (Panel A), reporting of math progress to parents (Panel B), and head teacher help in teaching grade 4 (Panel F). Notably, these improvements occur in all three study arms. Effect sizes are non-trivial: we observe a 15-20 percentage point increase in the likelihood of a student's 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 with teaching the grade 4 math class, for all three study arms. 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 in any study arm.¹² For head teacher attendance (Panel E), we observe differently signed time trends across arms (positive in the teacher treatment and control, negative in the head teacher treatment), although none of these changes is statistically significant.

We estimate Equation (1) for the same six outcomes, reporting results in Table 1, Panels B and C. As expected given Figure 2, we find little evidence of treatment effects. For the use of student groups, reporting math progress, and head teacher help, the null results reflect the similar (economically meaningful) improvements in outcomes under both treatment and control. For outside help and teacher attendance, the null results reflect the comparable lack of improvements in outcomes in treatment and control groups. There is a (weakly) significant negative treatment effect for head teacher attendance: the teacher treatment reduced head teacher attendance by six percentage points relative to the control (albeit from a high base). This result reflects the differently signed time trends in Figure 2, Panel E.

¹² There is an improvement in helping with outside classwork in the control group that is (weakly) statistically significant, but the magnitude is trivial.

5 Discussion

In this Section, we explore why there were similar improvements in outcomes relating to in-class teaching practice across *all* study schools. Centrally here, the government insisted on using E&SE officials for the school visits, so the same individuals were conducting teacher and head teacher visits in treatment and control schools. This design choice was made because it is these officials who visit schools, not other staff so that the new accountability system could be embedded within existing administrative structures. Perhaps inevitably, it appears that E&SE officials adopted similar behaviors in all schools, both treatment and control.

One such behavior is coaching. Under 'business as usual' visits to schools, E&SE officials played an active role, giving guidance and feedback to teachers. During the pilot, they were asked to continue this ('business as usual') practice in control schools but to switch to a passive monitoring role in treatment schools. Two sources of evidence suggest that they found this difficult and continued to coach.

First, we watched 357 videos of teachers during a classroom observation of teacher pedagogy. In one-quarter of all treatment schools (we obviously lacked such videos for control schools), we found documentary evidence that, rather than simply noting a score, the E&SE official intervened to correct a teacher's actions during what should have been a passive classroom observation. Common behaviors were asking the teacher to split children into groups and prompting teachers to use learning materials. These are both pedagogical techniques that can be helpful for learning. There were also instances in which the E&SE official referred to an interaction that occurred prior to the classroom observation and recording of the video—for example, the official can be heard on camera saying, "do X as I told you to do before." These videos suggest that some officials either could not or chose not to switch to a passive monitoring role. Coaching may also have occurred off-camera (or was difficult to assess and code), so these already large numbers are likely underestimates.

We also conducted detailed, semi-structured qualitative interviews with 6 of the 20 E&SE officials in the study. The interviews also indicate that officials continued to coach teachers on aspects of pedagogy. For example, one official stated that *"after the lesson was finished, 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, I would tell them."*

In addition to continuing to coach in treatment schools, it appears that some E&SE officials may have tried to take the accountability system into control schools. In one interview, an E&SE official noted that teachers in control schools were asking for advice more actively because some officials had told them that the pilot would eventually be rolled out in their schools as well. Even more tellingly, one official remarked, that *"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."*

This supplementary evidence that E&SE officials took coaching into treatment schools and (aspects of) accountability into control schools may explain why we saw in-class teaching practices—use of student groups, reporting math progress to parents, and head teacher help teaching grade 4—improve from baseline to endline across *all* study schools.¹³ There were no improvements in educator attendance, nor in helping students individually outside of school, however. We hypothesize that those outcomes may be higher-cost and lower-reward for educators. Splitting students into groups and keeping parents informed are potentially relatively easy for teachers to do; helping students outside of school is more costly. Teacher and head teacher attendance was already quite high, so further improvements may have been perceived as too costly.

One interpretation of these findings is that some E&SE officials found it difficult to change their 'mental model' and continued to feel that it was their responsibility to help and provide advice. The 'business as usual' protocol officials wrote prior to hearing about the interventions suggests that a lot of interaction and coaching occurred. During the interviews, one official expressed a desire for the space to coach: *"if we were able to choose from some of the other teachers that we knew were lacking or who were good, and we could make a video of them and show from the checklist that these are some of the inadequacies of your teaching. Those are some of the things we could not do."* After the pilot, the Government of KP began implementing a similar program called the School Quality Management Initiative (SQMI), and the same

¹³ On head teacher help, one inspector stated that, *"there were a lot of changes and the head teacher especially became more involved. Before she used to sit like an authority figure and say do this or do that, but after the project, they became more involved and took care of the teachers and school and the children's studying more."*

official said that, "*in the SQMI project that we are doing now, we have the option of giving feedback.*"¹⁴

The above discussion provides a plausible explanation for the absence of treatment effects on teacher behavior. But it also raises the question as to why improvements in teaching practices only resulted in small learning gains in all schools. Here we can only speculate, but we suspect that the returns to improved teaching practices take time to become apparent, and certainly longer than the (enforced) 5-month duration of this pilot.

6 Conclusion

This study has shown that changing bureaucrats' mental models can be difficult. Using existing government officials to undertake performance monitoring clearly has benefits, as accountability can be embedded within the system. However, staff who previously visited schools to provide advice and feedback may need encouragement to change their 'mental model' and to switch from this active role to that of a passive performance monitor. In our study, while some officials did appear to change their mental model, many did not. Stronger incentives—e.g., salient rewards for appropriate monitoring, and perhaps sanctions for poor monitoring—may be necessary to encourage all officials to make the switch, if that is the goal.

Whether governments *should* be pursuing an accountability system (with training and/or incentives for officials asked to monitor performance) remains an important question. Another policy would be to harness the apparent enthusiasm for coaching by supporting government officials to provide more systematic in-service training at the frontline. These policies are not mutually exclusive and governments could in principle do both. The key takeaway from this paper is that coaching and monitoring are strategically different and may need to be performed by different personnel.

¹⁴ The SQMI also measures test scores, attendance, and pedagogy (without photos or videos). The main innovation is that it allows inspectors to engage in a *structured* process to provide feedback and to coach on good practice, which had been lacking before.

References

- Abadie, A., & Athey, S., & Imbens, G. W., & Wooldridge, J. (2017). "When should you adjust standard errors for clustering?" (No. w24003). *National Bureau of Economic Research*.
- Anderson, M. 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.
- Bana, B, McCourt, W. (2006). "Institutions and Governance: public staff management in Tanzania." *Public Administration and Development*, 26: 395-407.
- Barlevy, G., & Neal, D. (2012). "Pay for Percentile." *American Economic Review*, 102(5), 1805-31.
- Cilliers, J., & Mbiti, I.M. & Zeitlin, A. (2020). Can public rankings improve school performance? Evidence from a nationwide reform in Tanzania. *Journal of Human Resources*, pp.0119-9969R1.
- Evans, D. K., & Popova, A. (2015). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. The World Bank
- Figlio, D. N., & Getzler, L. S. (2006). "Accountability, ability and disability: Gaming the system?" *Advances in Applied Microeconomics*, 14, 35-49.
- Figlio, D. N. & Winicki, J. (2005). "Food for thought: the effects of school accountability plans on school nutrition." *Journal of Public Economics*, 89(2), 381-394.
- Gilligan, D., & Karachiwalla, N., & Kasirye, I., & Lucas, A., & Neal, D. (2019). Educator Incentives and Educational Triage in Rural Primary Schools, *Journal of Human Resources*, 1118-9871R2.
- Glewwe, P., & Ilias, N., & Kremer, M. (2010). "Teacher incentives." *American Economic Journal: Applied Economics*, 2(3), 205-227.
- Glewwe, P., & Muralidharan, K. (2016). "Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications." In *Handbook of the Economics of Education* (Vol. 5, pp. 653-743). Elsevier.
- Habib, M. (2015). Teacher and School Administrator Incentives for Improved Education Delivery in Khyber Pakhtunkhwa Province, Pakistan. *The International Growth Center, Working paper*, May 2015.
- Haeck, C., & Verboven, F. (2012). "The internal economics of a university: Evidence from personnel data." *Journal of Labor Economics*, 30(3), 591-626.

- Hambleton, R.K., & Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Sage.
- Holmstrom, B. & Milgrom, P. (1991). Multi-Task Principal-Agent Analysis: Incentive Contracts, Asset Ownership, and Job Design, *Journal of Law, Economics, and Organization*, 7: 24-52.
- Muralidharan, K., & Das, J., & Holla, A., & Mohpal, A. (2014). The Fiscal Costs of Weak Governance: Evidence from Teacher Absence in India. *NBER Working Paper 20299*.
- Karachiwalla, N., & Park, A. (2017). "Promotion Incentives in the Public Sector: Evidence from Chinese Schools," *Journal of Public Economics*, 146, 109-128.
- McKenzie, D. (2012). "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics*, 99(2), 210-221.
- Neal, D. (2011). The design of performance pay in education. In Handbook of the Economics of Education (Vol. 4, pp. 495-550). Elsevier.
- Nieddu, M., & Pandolfi, L. (2018). The Effectiveness of Promotion Incentives for Public Employees: Evidence from Italian Academia (No. 507). Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy.
- Pakistan Public Administration Research Centre. (2004). A Guide to Performance Evaluation. Islamabad.
- Stallings, J. (1977). Learning to look: A handbook on classroom observation and teaching models. Wadsworth Publishing Company.

Table 1: Impact estimates of teacher and head teacher interventions

Panel A: Student Outcomes			
	Proportion Correct (1)	Student Present (2)	Dropped Out (4)
Head Teacher Treatment	-0.004 (0.008)	-0.030 (0.016)	0.002 (0.004)
Teacher Treatment	0.006 (0.009)	-0.029 (0.016)	-0.001 (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 (1)	Teacher Splits Students into Groups (2)	Teacher has Spoken to Parents about Progress (3)
Head Teacher Treatment	-0.009 (0.041)	-0.008 (0.043)	-0.039 (0.041)
Teacher Treatment	0.044 (0.044)	0.023 (0.043)	0.040 (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 (1)	Head teacher Present (2)	Head teacher Helps G4 Math (3)
Head Teacher Treatment	-0.001 (0.035)	-0.035 (0.024)	-0.061 (0.055)
Teacher Treatment	0.025 (0.031)	-0.061* (0.026)	-0.064 (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 school level 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 (Anderson, 2008). Panel A: proportion of questions answered correctly by grade 4 students on endline written grade 4 math assessment (of 45 questions); indicator variable for grade 4 student present at endline survey (measured by the enumerator); indicator variable for grade 4 student dropped out of school between the baseline to endline surveys (measured by the enumerator). Here, the baseline control variable 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 math progress. 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.

Figure 1: Student Outcomes

Notes: Outcomes are reported at the student level ($N = 4,639$ in Panel A, $6,389$ in Panel B, and $6,407$ in Panel C). 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.

Teacher Practices

Panel A

Proportion of Students Reporting:
that their teacher splits students into groups

Treatment Group	Baseline	Endline	p-value
Teacher	0.48	0.68	0.001
Headteacher	0.42	0.65	0.001
Control	0.42	0.62	0.001

Panel B

Proportion of Students Reporting:
their teacher gives their parents feedback on progress in math

Treatment Group	Baseline	Endline	p-value
Teacher	0.28	0.42	0.001
Headteacher	0.26	0.38	0.001
Control	0.26	0.36	0.001

Panel C

Proportion of Students Reporting:
that their teacher helps them outside the classroom

Treatment Group	Baseline	Endline	p-value
Teacher	0.62	0.60	0.836
Headteacher	0.60	0.58	0.303
Control	0.60	0.57	0.065

Educator Effort

Panel D

Teacher Present on the Day of School Visit
by treatment group

Treatment Group	Baseline	Endline	p-value
Teacher	0.96	0.98	0.652
Headteacher	0.92	0.95	0.362
Control	0.95	0.95	1.000

Panel E

Headteacher Present on the Day of School Visit
by treatment group

Treatment Group	Baseline	Endline	p-value
Teacher	0.97	0.94	0.471
Headteacher	0.95	0.97	0.714
Control	0.99	1.00	0.320

Panel F

Head Teacher Helps Teach Grade 4 Math
by treatment group

Treatment Group	Baseline	Endline	p-value
Teacher	0.68	0.82	0.043
Headteacher	0.62	0.81	0.005
Control	0.68	0.88	0.001

18

Online Appendix

Appendix A: Additional Details

A.1 Existing accountability systems

The KP government has two accountability systems in place that should, in theory, lead to high effort from bureaucrats, including educators. The first is the Performance Evaluation Report (PER). Teachers are evaluated annually using the PER, usually completed by head teachers, which feeds into promotion decisions. Additionally, a school inspection system exists whereby a district education official (who is independent of the school) should conduct quarterly school visits and report findings to the district education office. For teachers, performance evaluations should carry rewards and consequences for promotions (Pakistan Public Administration Research Centre, 2004). For schools, inspections should reward and punish head teachers, inducing greater effort from all teachers in their school, and improving learning outcomes. In practice, both systems are largely ineffective.

The PER system for teachers suffers three main problems. First, there is no teacher-specific PER -- only a general civil servant PER that does not directly measure aspects of teaching. It is very general and is extremely subjective).¹ As Holmstrom and Milgrom (1991) show, employees focus effort on measured and rewarded tasks, meaning that the present PER system does not incentivize good teaching. Second, PERs are conducted at an inappropriate time of year—at the end of the calendar year, so mid-way through the school year – making it unclear which teacher deserves the credit or blame for student outcomes. Finally, PERs are carried out by the headteacher, who -- in Pakistani society -- finds it difficult to criticize colleagues. This structure limits the quality of the information of PERs; all teachers tend to receive the same passing score (ratings compression) (Pakistan Public Administration Research Centre, 2004), rendering PERs meaningless.

The current system of school inspections is also highly problematic. In theory, district government officials should carry out inspections quarterly and issue warnings to headteachers not complying with policies and regulations. In practice, inspections are irregular. There is little guidance on what should occur. The inspectors observe and note

¹ The PER includes measurements such as personal qualities (for example, integrity), and the guidelines even stipulate the type of typewriter that should be used to write up the report.

schooling aspects that are not meaningful to learning (for example, students' cleanliness), and warnings to higher-ups are rarely provided. Inspections, when they occur, thus carry little meaning. This is unfortunate, as a study in India (Muralidharan et al., 2014) found that monitoring schools -- even without financial rewards -- had one of the strongest impacts on teacher attendance compared to other school inputs (including hiring more teachers). Further weakening both systems, neither PERs nor inspections are tied to rewards or sanctions in practice. Promotions are awarded based on tenure and connections. Thus, despite having two potential systems to hold teachers and schools accountable, they do not motivate high effort.

A.2 New accountability systems evaluated

During the 2017–18 school year, the research team, working closely with the KP E&SE Department, introduced two new accountability systems with the aim of motivating educator effort and improving student learning outcomes. Both interventions took the example of the Chinese promotion system for teachers in terms of which outcomes were measured (Karachiwalla and Park, 2017) and built on evidence on how to incentivize educators effectively (Cilliers et al., 2020; Barlevy and Neal, 2012; Gilligan et al., 2019). To make the intervention tractable and facilitate impact evaluation, we focused on head teachers and fourth-grade math teachers and their students. Fourth-grade students are old enough that they do not require tests to be administered orally and can answer questions about their school and home, but are also young enough that dropout rates are still a concern and interventions affecting their attendance, learning, and propensity to drop out provide useful policy recommendations.

The first intervention was the Annual Teacher Evaluation (ATE), which aimed to address the shortcomings of the PER in several ways. It was intended to cover the school rather than the calendar year, and so a single cohort of students. It focused on teaching-specific outcomes—presence of the teacher and his/her students, the teacher's pedagogy, and test scores of the teacher's students. It was conducted by relatively more independent, district-level officials rather than colleagues from the same school. Finally, and crucially, teachers' ATE scores were explicitly linked to career progression via promotion tournaments within school districts; teachers who performed the best relative to others in the same school district and treatment arm would have their promotions fast-tracked by a year, while those who performed worst relative to others in the same school district would have their promotions

delayed by a year (teachers in the middle would face no changes in promotion speed). We refer to this as the *teacher treatment* (T) since it pertained to teachers.

The second new accountability system was the School Inspection Report (SIR). This school-level report was also undertaken by independent district officials and, in contrast to the previous school inspection system, was based on regular, structured site visits. Similar information is collected as for ATEs—presence of the head teacher, all teaching staff and students within the school, and teacher pedagogy. A key difference compared to the ATE is that it was the head teacher's career progression that was to be affected; head teachers whose schools performed the best relative to schools in their school district in the same treatment arm would have their promotions fast-tracked by a year, while those who performed worst relative to other schools would have their promotions slowed by a year (head teachers in the middle would face no changes in promotion speed). We refer to this as the *head teacher treatment* (HT) since it is the head teacher whose career was affected.

In October 2017, 41 district inspectors were trained for three days on how to carry out both the teacher and headteacher interventions. These included the following officials: the District Education Officer (DEO), the Deputy District Education Officer (DDEO), the Sub-District Education Officer (SDEO), and the Assistant Sub-District Education Officer (ASDEO). The DEO and DDEO work at the level of the district (this is the geographic district, one administrative unit below the province), the SDEO works at the level of the *Tehsil* (the administrative unit below the district), and the ASDEO works at the level of the circle. Circles are equivalent to school districts in the United States. The circles are separated by gender. (Schools in Pakistan are gender-segregated, so there are male schools, circles, DEOs, and female schools, circles, and DEOs.) The ASDEO is the main point of contact between the Education Department and schools. Before introductions to the study, these district officials created documents describing their current practices for school visits at the very start of the training. This resulted in a "business as usual protocol." During the training, inspectors were taught about randomization and the importance of adherence to the treatment group protocols (including the business as usual protocol in control schools).

The details of the proposed teacher intervention are as follows. All schools were scheduled to receive five unannounced visits from an independent district official during the 2017-18 school year; ultimately, 2-3 were held due to logistical complications. The district officials involved were the DEOs, DDEOs, SDEOs, and ASDEOs. District officials would only carry out the school visits in the circle/Tehsil/district to which they belonged. The order of school

visits was randomly assigned by the research team, effectively making the date of each inspection random. All district officials were provided with: the teacher treatment protocol; a tablet (pre-loaded with a relevant computer-assisted personal interview (CAPI) program with which the intervention was conducted, including teacher/student rosters from the research team's baseline survey data collection) and a tripod stand for video recordings; a paper teacher pedagogy assessment tool to use during classroom observations; and a Wi-Fi box with credit to upload data. Upon arrival at the start of the school day for the first visit, the district official explained the new performance evaluation system to the grade 4 math teacher and found out when the Grade 4 math class would be taught. Each district official would then collect information on four performance measures: the presence and pedagogy of the Grade 4 math teacher and the presence of and learning outcomes for his/her students in the Grade 4 math class.

The scoring rubric was outlined as follows. These four performance measures were scored for each visit. For the first three measures, scoring was based on absolute performance against a pre-specified objective rubric.

- Teacher attendance: present or excused absence = 8 points, unexcused absence = 0 points. This was measured by taking a photo on the tablet of either the teacher with his/her national ID card or a photo of proof of excused absence (for example, a letter asking the teacher to attend training or an SMS from the head teacher giving the teacher permission to be absent).
- Student attendance: sliding scale from 0 to 8 points for the share of students present (max for >90 present). This was measured by taking each student's attendance on the research team's baseline survey roster and entering their attendance on the tablet. The tablet then calculated the share of students present. Finally, a group photo of the students present was taken.
- Teacher pedagogy: sliding scale from 0 to 8 points for the average share of students engaged in active teaching activities (max for >88 percent engaged). This was measured using an observation tool developed by the research team. The district official would record a video of the class while filling out the observation tool on paper; this required the district official to record up to three activities that were occurring during the first ten seconds of each five-minute interval, as well as, for each activity, the number of students that were actively engaged in each activity. After the class, the district official would enter the data into the tablet, and the tablet would calculate a score over the 5-minute intervals for which class was in session.

For the last measure, student learning, the inspector asked students in the class 50 oral math questions. Students were lined up in alphabetical order and cycled through

multiple times if necessary until 50 questions were asked. Scoring was based on relative performance and was intended to capture the main features of the 'pay-for-percentile' approach (Barlevy and Neal 2012). Specifically, once all data were collected from all 80 teacher treatment schools, Grade 4 math teachers were placed into 'bins' of 8 based on the proportion of questions answered correctly by their students at the start of year (first visit). At the end of the year (last visit), these teachers' students were ranked within baseline score bins. The Grade 4 math teacher in the circle with the highest average rank across bins would receive eight marks, the next highest average rank seven marks, down to zero for the lowest average rank. The research team would perform this scoring of student learning.

The scores across these four performance measures on each visit were aggregated into a single, cardinal score for each Grade 4 math teacher. Specifically, the overall score was the sum of all measures for all visits (teacher attendance $\times$ # visits, student attendance $\times$ # visits, teacher pedagogy $\times$ # visits, student performance $\times$ # visits) scored on the 0-8-point scale (while there was a single measure of student performance, it would be included multiple times—equal to the number of visits—in the score calculation to ensure that the four categories were weighted equally as teachers were informed). These cardinal scores fed into an ordinal ranking exercise. Within each of the 20 circles in the study, there were four schools randomly assigned to the teacher treatment. At the end of the school year, the four Grade 4 math teachers within each circle teacher treatment pair were ranked based on their cardinal score. In the event of a tie, the district officials would together consider all four dimensions of performance and break the tie in favor of one teacher or the other. The top-ranked teacher would have his/her promotion fast-tracked by one-year, while the bottom-ranked teacher would have his/her promotion delayed by one-year. The two teachers in the middle would experience no change in the speed of their promotion.

To reduce behaviors that improved teachers' scores but not student learning, not all details of the scoring process were communicated to teachers (e.g., which pedagogical activities were considered active, and the math test questions). The detailed protocol for all four dimensions of this scoring rubric was communicated to inspectors during their training and to the Grade 4 math teacher at the start of the first surprise visit, via an information sheet outlining the intervention and the scoring.

The head teacher intervention was identical to the teacher intervention except that each inspector was to collect information on the head teacher's presence, all teachers and students in the school, and conduct a classroom observation of two teachers. The scoring

rubric was as follows. These four performance measures were summed to produce an overall score for each visit:

- Head teacher attendance: present or excused absence = 8 points, unexcused absence = 0 points. This was measured by taking a photo on the tablet of either the head teacher with his/her ID card or a photo of proof of excused absence.
- Teacher attendance: sliding scale from 0 to 8 points for the share of teachers present (max for >90 percent present). This was measured by taking a photo of each teacher in the school either with his/her national ID card or a picture of proof of excused absence. The tablet automatically calculated the score.
- Student attendance: sliding scale from 0 to 8 points for the share of students present (max for >90 percent present). This was measured by taking a photo of each class and entering the number of students present.
- Teacher pedagogy: sliding scale from 0 to 8 points for the average share of students engaged in active teaching activities (max for >88 percent engaged) across two classroom observations. This was measured using the same process as for the teacher treatment but averaging over the two classes observed.

The detailed protocol for all four dimensions of this scoring rubric was again communicated to the district officials during training and to the head teacher at the start of the first surprise visit via an information sheet. The scores across these four performance measures overall visits were aggregated into a single, cardinal score for each head teacher; specifically, the overall score would be the unweighted sum of all measures for all visits (head teacher attendance $\times$ # visits, teacher attendance $\times$ # visits, student attendance $\times$ # visits, teacher pedagogy $\times$ # visits) scored on the 0-8-point scale.

As in the teacher intervention, these cardinal scores were fed into the ordinal ranking exercise within circles. At the end of the school year, the schools within each circle/ treatment pair were ranked based on their cardinal score, and ties were broken as for teachers. The head teacher of the top-ranked school would have his/her promotion fast-tracked by one-year, the head teacher of the bottom-ranked school would have his/her promotion delayed by one-year, and the two head teachers of middle-ranked schools would experience no change.

The interventions came close to achieving the three school visits planned as of the start of the intervention period (though fewer than the originally envisioned five visits). Overall, 96 percent of treatment schools received visit 1, 46 percent received visit 2, and 98 percent received visit 3. Visit two was halted after having been completed in only a subset of treatment schools to ensure that there was enough time (given the amount of time visits were taking to date) to visit every school as part of visit 3. This resulted in an average of 2.7 visits

in the teacher treatment arm (on average lasting 7.1 hours) and 2.5 visits in the head teacher treatment arm (on average lasting 6.2 hours). As schools were randomly assigned visit dates, the total number of visits received is not correlated with unobserved school characteristics. Nevertheless, we omit visit two from any analysis. Unfortunately, we did not obtain data on the length of 'business as usual' inspections in control schools. Still, the completion of a 'business as usual' protocol by district officials in advance of explaining the interventions indicated considerably shorter and lighter touch visits in control schools: on average, under 50 minutes. Most school visits were performed by ASDEOs, with only 18 percent performed by DEOs or DDEOs in each of the treatment arms in visit 2. Visits 1 and 3 were conducted exclusively by ASDEOs.

To incentivize district officials to undertake the ATE and SIR thoroughly and objectively, the scores submitted were to be audited by individuals with an arms-length relationship to the district officials/schools: either a district official (DEOs, DDEOs, SDEOs, ASDEOs) from non-study districts or a secondary school teacher from study districts. Auditors were to review information collected during the visits (photographic proof of attendance, video classroom observations) and note several aspects of the visit (e.g., was there missing data, was the video camera aimed so that the teacher and all students could be seen, were photos of teachers with IDs taken correctly, were appropriate forms of excused absences noted). District officials performing the school visits were to have their promotions also fast-tracked, maintained, or delayed based on the audits of how well they performed the visits. However, the government only agreed to put a note in their promotion dossiers.

A.3 Sampling details

We identified three districts in which it would be both logistically and politically feasible to operate in primary schools: Charsadda, Mardan, and Nowshera. There were 48 circles (24 male circles and 24 female circles) in these three districts during the fall of 2016 (September – November)²—the period for which we had administrative data from which to draw our sample of schools. We first compiled a list of all primary schools in our three districts; there

² Our focus on these three months alone and not on earlier months in 2016 is motivated by the timing of enrollment drives and time needed for enrollments to stabilize. The school year starts in May when there is an initial enrollment drive, but this is mostly a period of textbook distribution, and schools then break for June – August. There is then a second enrollment drive and schools restart in September. We chose to focus on September – November as enrollment should have stabilized by then.

were 3,111 schools³. Given our interest in grade 4, our next concern was having sufficient students in grade 4 to implement an intervention. We thus eliminated schools with fewer than ten or more than 50 students in grade 4⁴. We were therefore left with 1,877 schools. We also eliminated schools not classified as public schools and urban schools (where accountability tends to be stronger), leaving a sample of 1,634 schools⁵.

For the interventions to be relevant, we also needed additional information to draw our sample. We received data from the E&SE Department on which of these schools had a grade 4 teacher or headteacher who was up for a promotion or transfer in the next two years (we eliminated such schools as educators would be resistant to interference); whether the grade 4 math teacher was also the headteacher (the incentive could be provided to one or the other but not both, so we eliminated schools where this was the case); and whether the Pakistan Reading Programme (PRP)⁶ was in place (we eliminated such schools as they would not be comparable to the average school). This resulted in a total sampling frame of 870 schools from which to draw our sample. Given that we needed 12 schools in each circle (4 per treatment arm), we needed to draw from circles that had sufficient schools that could serve as replacement schools should the survey team uncover that a school was closed, or for other reasons could not be included in the study (for example, the grade 4 teacher actually was the head teacher). There were 15 male circles and five female circles that met this criterion.

We randomly ranked schools within each circle-gender, and the survey team visited them in order. They confirmed the information set we already had to certify school eligibility. If a school ended up being ineligible, it was excluded, and they continued down the list until they had 12 eligible schools in each circle—resulting in our sample of 240 schools.

³ The list of schools were those that had non-missing enrollment data in the Independent Monitoring Unit's (IMU) database for at least one month during September, October, and November 2016.

⁴ We were further concerned about schools with too many students in grade 4. Such schools may have multiple grade 4 teachers, making incentives much more complicated. Further, we were constrained by our inability to administer tests to a very large pool of students.

⁵ Of the 1,877 schools we dropped 60 schools not classified as public schools (bringing the total to 1,817 schools), one school that was oddly coded as a secondary school despite being in the database given to us (bringing the total to 1,816), and 181 urban schools (bringing the total to 1,635). Finally, we dropped one school for which information on its circle was missing (bringing the total to 1,634 schools).

⁶ The PRP is a USAID funded project that targeted many schools in KP province and included training on teaching students how to read and also provided learning materials.

A.4 Treatment assignment and power calculations

Following the baseline survey, we used a re-randomization method to randomly allocate four schools within each circle to each of three groups: teacher treatment, headteacher treatment, and control. We selected 18 school, teacher, headteacher, and student variables to test balance and performed 1,000 replications of random assignment within circles. We then selected the replication that provided the minimum-maximum t-statistic (the replication for which the highest t-statistic was the lowest).

Before the study, power calculations were performed using data from the Learning and Educational Achievements in the Pakistan Schools (LEAPS) study conducted in Punjab, Pakistan. No test score data from KP was available. Additionally, the LEAPS study tested students in grade 4 math—the same level and subject targeted by this study. We used an alpha of 0.05, power of 0.8, and assumed one round of follow-up data collection.⁷ With 240 schools (80 in each treatment arm) and an average of 20 students per school (our target), we were powered to detect an effect size of 0.1 standard deviations in test scores. An effect size of 0.2 standard deviations difference in test scores is considered a typical effect size in the education literature (see Evans and Popova, 2015). Since test scores were our primary outcome variable of interest, power calculations were only performed on this outcome.

A.5 Details on test construction

The baseline and endline survey tests were based on the Pakistani math curriculum for grades 1-3. We created test questions for each grade level and piloted 60 questions in 18 schools. We then used Item Response Theory (Hambleton, Swaminathan, and Rogers, 1991) to determine the proportion of students answering each question correctly and calculate the discrimination parameter (i.e., the degree to which the question can discriminate between students of differing ability). The items that almost none or all students answered correctly were removed, as were items with the lowest discrimination parameters. The baseline and endline exams had 45 questions each.

One component of the teacher treatment is math test scores. We began with 'item templates' for these oral assessments, which measured 50 specific 'constructs' (for example, two-digit addition, one-digit multiplication, etc.). The constructs were again part of the

⁷ We computed and used a standard deviation of 1.121 for math test scores, a correlation between baseline and endline test scores of 0.665, and an intra-cluster correlation within schools of 0.0417.

grades 1-3 Pakistani math curriculum. The piloting of the 50 templates took place in May 2017 in 32 schools. We used the proportion correct to drop questions that either everybody or nobody got correct. We then constructed four questions from each template (for each construct) to build an item bank of 200 questions that could be randomly drawn by the tablet to administer in each school for the teacher intervention so that no two schools would get the same set of questions. Questions were drawn at random from *within* constructs to ensure that all schools were tested on the same material and that tests were not different in terms of difficulty.

Figure A1: Timeline of Events

Table A1: Student-level summary statistics and balance across study arms

Variable	Study Arm						Pair-wise Differences		
	(1) Control	(2) N	(3) Head Teacher	(4) N	(5) G4 Teacher	(6) N	(7) C vs. HT	(8) C vs. G4 T	(9) HT vs. G4 T
Student gender - female	0.32	2,163	0.30	2,036	0.33	1,917	0.930	0.253	0.646
Student age (years)	10.51	2,031	10.47	1,884	10.50	1,766	0.777	0.812	0.739
Student was present on previous school day	0.87	2,163	0.87	2,036	0.87	1,917	0.754	0.619	0.732
Student was present today	0.84	2,163	0.81	2,036	0.82	1,917	0.165	0.184	0.893
Standardised Baseline IRT Score	0.04	1,807	0.01	1,655	-0.02	1,570	0.612	0.402	0.373
Household Asset Index	0.03	1,805	-0.06	1,655	0.02	1,571	0.223	0.780	0.648
Total siblings	4.82	1,805	4.91	1,655	4.88	1,571	0.384	0.459	0.727
Always enjoys going to school	0.67	1,805	0.69	1,655	0.62	1,571	0.286	0.306	0.038
Does school homework at home	0.92	1,805	0.90	1,655	0.90	1,571	0.332	0.630	0.734
Somebody helps with homework at home	0.72	1,657	0.74	1,492	0.69	1,419	0.432	0.310	0.112
Receive private tutoring for math	0.17	1,805	0.14	1,655	0.16	1,571	0.154	0.664	0.517
Puts hand up in most or every math lesson to answer questions	0.41	1,805	0.41	1,655	0.45	1,571	0.838	0.242	0.455
Teacher provided feedback to parents about math progress	0.28	1,805	0.27	1,655	0.29	1,571	0.931	0.609	0.728

Notes: Values in 'Pair-wise Difference' columns denote p-values. P-values were calculated from separate regressions of the baseline variable on each of the two treatment arms, including circle strata dummies, and standard errors clustered at the school level. The endline sample of schools is used as in the head teacher arm one school merged with a school outside the sample and was dropped; we focus on the analysis sample.

Table A2: School-level summary statistics and balance across study arms

Variable	Study Arm						Pair-wise Differences		
	(1) Control	(2) N	(3) Head Teacher	(4) N	(5) G4 Teacher	(6) N	(7) C vs. HT	(8) C vs. G4 T	(9) HT vs. G4 T
Head Teacher age	48.77	80	50.05	79	48.91	80	0.122	0.860	0.162
Grade 4 Teacher age	36.14	80	36.30	79	35.19	80	0.950	0.477	0.429
Head Teacher Years of Experience	7.81	80	8.95	79	7.67	80	0.420	0.920	0.386
Grade 4 Teacher Years of Experience	5.63	80	4.29	79	5.08	80	0.193	0.621	0.456
Head Teacher has obtained at least a B.A qualification	0.75	80	0.82	79	0.70	80	0.273	0.471	0.074
Grade 4 Teacher has obtained at least a B.A qualification	0.33	80	0.33	79	0.28	80	0.922	0.465	0.444
Total number of teachers in this school	4.25	80	4.43	79	4.06	80	0.520	0.450	0.135
Total number of students enrolled in this school	177.40	80	186.16	79	171.51	80	0.419	0.561	0.162
Student-teacher ratio	43.63	80	43.94	79	42.89	80	0.898	0.671	0.607
School is supported by an NGO	0.17	80	0.23	79	0.26	80	0.396	0.160	0.601
Teachers received training last year to improve subject knowledge	0.80	80	0.78	79	0.74	80	0.830	0.313	0.419
Teachers received training last year to improve teaching method	0.80	80	0.78	79	0.81	80	0.815	0.831	0.667
School has clean drinking water available	0.90	80	0.92	79	0.85	80	0.594	0.351	0.148
School has access to electricity	0.91	80	0.90	79	0.94	80	0.753	0.556	0.397
Index of school resources	0.00	80	0.07	79	-0.09	80	0.723	0.653	0.460

Notes: Values in 'Pair-wise Difference' columns denote p-values. P-values were calculated from separate regressions of the baseline variable on each of the two treatment arms, including circle strata dummies, and standard errors clustered at the school level. The endline sample of schools is used as in the head teacher arm one school merged with a school outside the sample and was dropped; we focus on the analysis sample.

Table A3: Attrition

	(1) Student	(2) Grade 4 Teacher	(3) Head Teacher
Teacher Treatment Arm	0.002 (0.005)	0.038 (0.028)	0.100** (0.050)
Head Teacher Treatment Arm	0.012 (0.012)	0.000 (0.018)	0.013 (0.047)
Observations	6493	240	240
R^2	0.037	0.073	0.186
Mean Dependent Variable	0.013	0.025	0.125

*Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Results derived from regressions of a dummy variable for a student, teacher, or head teacher being present for the survey at endline given that they were present at baseline. Strata dummy variables for circles (the level of sampling stratification) are included, and standard errors are clustered at the school level.*

Appendix B: Business as Usual Protocols

Table B1: Business as Usual Inspection Visits

Question	Charsadda female district officials	Charsadda male district officials	Mardan female district officials	Mardan male district officials	Nowshera male district officials
1-Average # of visits per school per year	5-6	Once every month	4	3-6 times per year	5
2-Activities in last inspection	<ul style="list-style-type: none"> -Checking attendance registers -Checking discipline and cleanliness of students -Checking classes -Writing in school's logbook -Student evaluation -Parent-Teacher Council (PTC) register/record checked 	<ul style="list-style-type: none"> -Physical checking of conditional grant utilization -PTC utilization -Academics of students -Teacher attendance -School register and records maintenance -Checking access to basic facilities 	<ul style="list-style-type: none"> -Teacher and student attendance -Conditions of the building and needs -Student dropout and its reasons -Fund utilization 	<ul style="list-style-type: none"> -Attendance of the school staff -Go to classes and ask students questions -Try to solve teachers' problems -Help the teachers understand teaching methods -Checking of school records 	<ul style="list-style-type: none"> -Teacher attendance -Student attendance -Monitoring of grants work -Children's education/work -Attendance of security guard -Condition of the building -Assessment of school needs and problems faced by teachers
3-Frequency of visits (how many school visits personally by district official)	30-35 schools	Visit schools every day	25-30	15 visits a month	2-4
4-Inspection length in minutes	30-50	40-60	60	60	40-60
5-Method for deciding which schools to visit and	A tentative monthly plan is made by the head of the department, and	We decide ourselves about which schools to visit and when, or	As per approved tentative program by department and sometimes on	We make the schedule and share it with the DEO and SDEO for approval.	DEO decides, or I decide myself and get approval by the DEO and follow the schedule.

Question	Charsadda female district officials	Charsadda male district officials	Mardan female district officials	Mardan male district officials	Nowshera male district officials
when and how communicated	then the schedule submitted to the DEO office is followed.	it could be based on the request of the seniors, and sometimes we have a visit schedule that is provided in advance.	directive of higher-ups and then the tentative plan is approved by the DEO or SDEO.	Communication is by letter or phone.	
6-Degree of flexibility in timing of visit	Yes, sometimes.	Yes, it is flexible.	There is no flexibility.	No but we are given flexibility in office work during inspection days.	Yes, it is flexible.
7-Visits announced?	No	No	No	Yes, we inform the school for the annual inspection and conduct surprise visits for the rest of the inspection visits.	The teacher or head teacher is informed during the annual inspection and no one is told during the surprise inspections.
8-What happens if you cannot make it?	The DEO office is informed about it and the school is visited the next day.	We either send one of the juniors to conduct the inspection or sometimes reschedule it to the next day.	Visit the school the next day.	Either go on another day or someone else goes.	Senior staff then sometimes conduct the visit.
9-Inspection start time	8.45 a.m	About 8.30 a.m.	9 a.m. but also depends on distance of school from office.	About 8.30- 9 a.m. at morning assembly time.	Mostly before the morning school assembly, sometimes during the middle of the day.

Question	Charsadda female district officials	Charsadda male district officials	Mardan female district officials	Mardan male district officials	Nowshera male district officials
10-Tasks (in order performed and with length of time indicated in minutes)	<ul style="list-style-type: none"> -Check school attendance registers and teacher and class performance (5-10 min) -Check discipline and cleanliness of students (10-15) -Student registers and uniforms (10 min) -All other registers are checked (15 min) -Student evaluation in different classes (20-25 min) -Instruction and guidance for teachers (10 min) 	<ul style="list-style-type: none"> -Check attendance of teachers and registers and rest of staff (5-10 min) -Check utilisation of conditional grant and checking of classes (15-20 min) -Check the PTC record -Check basic necessities of the school, children's cleanliness, class work, general cleanliness of premises, homework completed (20 min) 	<ul style="list-style-type: none"> -Check student and teacher attendance (10 min) -Check curriculum and classroom activities and the time depends on subject and size of the class -Check the condition of the buildings -Check fund record -Assessing needs of schools -Solving difficult issues e.g., staff issues and complaints (time not indicated) 	<ul style="list-style-type: none"> -We first check the assembly (10 min) -Then check teacher attendance through records (2-3 min) -Then go to different classes and observe the teacher while he is teaching and also assess the performance of the students (45 min) 	<ul style="list-style-type: none"> -First check security guard attendance and teacher attendance record (5 min and depends on the number of teachers in the school) -Cleanliness of the school (10 min) -Check all the classes and sometimes visit the classes and ask questions from the students (time depends on class size) -Check availability of basic facilities as well as condition of the building (5-10 min) -Discussion of problems faced by teachers and students -Any community related issues -Cleanliness, enrolment (time taken varies)
11-Description of interactions with head teachers and teachers	Interaction is mostly with the head teacher but also teachers and other staff and all school related	Meet with the teachers to discuss teaching methods and with head teachers to discuss discipline of schools,	Meet with all the teachers and discuss school and student improvement mechanisms and discuss cleanliness	Yes, meet with all the teachers and discuss the main problems and challenges that teachers are facing when teaching.	Discuss teacher performance and teaching style with teachers and with head teachers attitudes and performance of each class teacher, whether

Question	Charsadda female district officials	Charsadda male district officials	Mardan female district officials	Mardan male district officials	Nowshera male district officials
	issues are discussed.	cleanliness of school as well as teaching and learning skills.	and discipline with the head teacher.	With the head teacher discuss the supervision of the teacher.	they are competent and are giving special attention to the children.
12-Description of interactions with students	Interact with students of all grades in connection with learning problems, uniforms, regularity, punctuality and results.	Yes, I meet with the children and discuss their problems as well as challenges faced by them. I meet with grade 5 children who are about to appear for board exams.	We meet with students and check an activity based on learning.	Meet with students of all grades and discuss with them the issues related to their schooling.	Go and meet children in all the classes and talk to them about games as well as what they are learning.
13-List of information recorded during inspections	<ul style="list-style-type: none"> -Teaching and non-teaching staff presence -Discipline and cleanliness -Enrolment and absenteeism -Register work and if it is good or not -Utilisation of funds and physical verification -School security issues -PTC record checking 	<ul style="list-style-type: none"> -School discipline -Cleanliness -Painting the school walls -Teaching and learning activities -Monitoring conditional grant activities -Teacher attendance, notebook/homework checking -Student attendance -Main facilities 	<ul style="list-style-type: none"> -Student teacher ratio of school -Dropout and enrolment of students -School needs assessment -Community issues 	<ul style="list-style-type: none"> -Staff attendance -Making sure the class is progressing in line with the syllabus -Making a list of all the needs of the school 	<ul style="list-style-type: none"> -Is the security guard available? -Teacher attendance -Student attendance -Head teacher and teacher duties and problems faced by them -Education of children

Question	Charsadda female district officials	Charsadda male district officials	Mardan female district officials	Mardan male district officials	Nowshera male district officials
14-Who do you give information to, how, and when?	Logbook copy in which visit details are recorded are submitted to department office. Monthly reports are also submitted to seniors. The information is stored in computers, files and diary.	Report details of visits to seniors, particularly DEO and SDEO and the information is stored in diary notebooks and logbooks.	We deliver all this information to the SDEO or DEO and it is delivered by hand. It is stored in files in the office.	Information is submitted to seniors in the form of a report. The main copy of the report is kept in the office in files and a photocopy of the report stays with the district official.	We note it in diary and logbook and the visit report is sent off to the seniors and the information is stored in hard copy files as well as in computerised documents.

Notes: These notes were recorded on paper during the initial training of district officials on the intervention, before anything was mentioned about the project. Text was entered as is and was only corrected for spelling and grammar.

ALL IFPRI DISCUSSION PAPERS

All discussion papers are available [here](#)

They can be downloaded free of charge

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

www.ifpri.org

IFPRI HEADQUARTERS

1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org