Title

Improving Accountability in Khyber Pakhtunkhwa’s Schools (IKAS)

Locations

Country

Pakistan

Region

Khyber Pakhtunkhwa province

Additional Trial Information

Keywords

Education

Additional Keywords

Teacher incentives, promotions, public service delivery, primary school

JEL Codes

I21, I24, O15

Abstract

Pakistan has low student learning levels and educator motivation, and accountability is a pervasive problem (Habib et al, 2015). Khyber Pakhtunkhwa (KP) province’s Elementary and Secondary Education (E&SE) Department wants to improve student learning and educator motivation by introducing two new accountability systems: one which links teacher promotions to their and their students’ performance, and one which links head teacher promotions to their and their school’s performance. While the province already has two accountability systems (teacher performance evaluations and school inspections visits) that should, in theory, motivate effort from teachers and from head teachers, respectively, both suffer from substantial shortcomings, rendering them ineffective as a means of motivating effort and providing accountability. They are irregular, uninformative, and do not clearly affect teachers or head teachers’ careers or salaries. This project will compare outcomes in 80 schools receiving the new teacher performance evaluation, 80 schools receiving the new school inspections, and 80 control schools. We seek to determine whether these interventions are effective tools for improving teacher and head teacher effort and improving student learning, and if so, which is most effective and cost effective.

Interventions

During the 2017/8 school year, the research team, alongside the KP E&SE Department, is introducing two new accountability systems with the aim of motivating educator effort and improving student learning outcomes. The first is the *Annual Teacher Evaluation* (ATE); under it, inspectors assign a score to teachers based on their own attendance, their students’ attendance, their pedagogy, and their students’ test scores. It improves on an existing but dysfunctional teacher evaluation system in a number of ways. It covers the school year rather than calendar year, and so a single cohort of students. It focuses 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 is conducted by relatively more independent, district-level inspectors rather than colleagues from the same school. Finally, and crucially, teachers' ATE scores are explicitly linked to career progression via promotion tournaments within districts; teachers who perform well relative to peers in similar schools are fast-tracked for promotion, while those who perform poorly are held back. We will focus on Grade 4 Math teachers since this is a subject for which language and translation are less likely to pose problems than for other subjects and where grading is especially objective, and also because students of this age can be interviewed alone and at low cost but are still relatively young.

The second new accountability system is the *School Inspection Report* (SIR). This school-level report is also undertaken by independent district inspectors. It contrasts with an existing but dysfunctional system of irregular and unstructured inspector visits. Similar information is collected as for ATEs, but focused on the full school rather than a particular teacher: presence of the head teacher, staff, and students within the school, and the pedagogy of randomly-selected teachers within the school. The main difference compared to the ATE is that it is the head teacher's career progression that is on the line; head teachers of schools that perform favorably compared to other schools in their circle (the equivalent of school district) are fast-tracked for promotion, while head teachers of schools that perform relatively poorly are held back.

The study sample of 240 schools consists of 20 circles (5 girls' circles and 15 boys' circles) with 12 schools in each circle, across three districts. Treatment was randomly assigned **within circles.** In each circle, there are 4 schools in the ATE treatment, 4 in the SIR treatment, and 4 in the 'business-as-usual' control.

ATE Intervention (80 schools)

All schools will receive three 'surprise' visits from an independent district inspector during the 2017/8 school year. The matching of inspectors to schools, and the dates of the site visits will be determined randomly by researchers. Inspectors will be provided the name of the school they are to visit the evening prior, and will be instructed not to notify the school. If the inspector cannot visit that school that day, he/she will need to provide a reason and, if this is acceptable, he/she will be reassigned an alternative, randomly selected school the following day. All inspectors will be provided with: a documenting outlining the ATE protocol; a tablet (pre-loaded with a relevant computer assisted personal interview (CAPI) program with which the ATE will be conducted, including teacher/student rosters) and tripod stand (to record classroom observations and student learning); a paper teacher pedagogy assessment tool; and a Wi-Fi box with credit to upload data. Upon arrival at the start of the school day on the first visit, the inspector will explain the new ATE system to the head teacher and find out when the Grade 4 math class will be taught. Each inspector will then provide an information sheet to the Grade 4 math teacher explaining the new ATE system and how it will affect that teacher's promotion and collect information on four performance measures: the presence and pedagogy of the Grade 4 math teacher, together with the presence of and learning outcomes (on 50 questions to be asked orally) for his/her students in the Grade 4 math class.

Scoring rubric. These four performance measures will be scored to produce an overall ATE score. For the first three, scoring is based on *absolute* performance on the following pre-specified and objective rubric.

- Teacher attendance: present or excused absent = 8 points, unexcused absent = 0 points.
- Student attendance: sliding scale from 0 to 8 points for share of students present (max for >90% present).
- Teacher pedagogy: sliding scale from 0 to 8 points for average share of students engaged in active teaching activities throughout the first 30 minutes of the class (max for >88% engaged).

For student learning, scoring is based on *relative* performance and is intended to capture the main features of the 'pay-for-percentile' approach (Barlevy and Neal 2012). Specifically, once all data have been collected from all 80 ATE study schools, Grade 4 math teachers will be put into 'bins' of 8 based on their start of year (first visit) 'percentage of 50 Grade 4 math questions answered correctly' score. This will create 10 separate tournaments of 8 teachers each. At the end of the year (third visit), teachers will be **ranked within bins**. The Grade 4 math teacher in the school with the top rank in his/her bin will receive 8 marks, the next rank 7 marks, down to zero for the teacher in the school with the bottom rank. This grouping of schools into bins and scoring of student learning within bins will be done by the research team.

The detailed protocol for all four dimensions of this scoring rubric will be communicated to inspectors during an October 2017 training (following a September 2017 baseline survey, but before treatment begins), and to the Grade 4 math teacher at the start of the first surprise visit, via a letter ("information sheet") outlining the intervention, number of visits, areas inspectors will score (but not specifics on how the four components will be scored), and how their overall score will be computed.

Career incentives. The scores across these four performance measures will be aggregated into a single, cardinal score for each Grade 4 math teacher. Specifically, the overall score will be computed using 10 measures (teacher attendance x3 inspections, student attendance x3, teacher pedagogy x3, and student performance), each scored on a 0-8-point scale. The teacher's final score will be the sum of the first nine measures plus three times the single, student performance measure. The measure of student performance will be included three times to ensure that the four categories are weighted equally, as the information sheet suggests. These cardinal scores will feed into an ordinal ranking exercise. Within each of the 20 circles in the study, there are 4 schools randomly assigned to the ATE treatment. At the end of the school year, the four Grade 4 math teachers x ATE treatment pair will be ranked on the basis of their *cardinal* ATE score. In the event of a tie, the inspector (who will have visited all schools in the comparison set) will be asked to consider all four dimensions of performance and break the tie in favor of one teacher or the other. The top-ranked teacher will have his/her promotion fast-tracked by one-year, while the bottom-ranked teacher will have his/her promotion delayed by one-year. The two teachers in the middle of the ranking will experience no change.

In an attempt to prevent dysfunctional behavior (collusion and/or demotivation) among teachers, teachers will receive basic details (in the information sheet) about the impacts of the inspections on their promotions (i.e. the possibilities of acceleration and delay based on relative performance), but not the precise details about the computation of ordinal rankings and comparisons of these with other teachers in the same circle and treatment arm.

Audits. To incentivize inspectors to undertake the ATE thoroughly and objectively, the scores submitted will be audited by individuals with an arms-length relationship to the inspectors/schools. There are two auditor treatments: in the first, auditors are drawn from a pool of *district officials* from non-study districts, and in the second, auditors are drawn from a pool of *secondary school teachers* from study districts. Each inspector will have one of their 'surprise visits' assigned to the district official treatment, and another to the secondary school teacher treatment. Both types of auditor will review all of the information collected during the visits (photographic proof of teacher and student attendance, video classroom observations) and will assign their own ATE score. Discrepancies between inspector and auditor scores will be noted. During training, inspectors will be told that their inspections will be audited and that their performance will be recognized and rewarded in two ways. First, inspectors who have performed well will be awarded a certificate at a public ceremony at the end of the school year. Second, inspector performance will be included in the dossier that forms part of their own Annual Confidential Report, and could therefore influence decisions about their own promotion and salary increments.

Marks assigned by the auditors will also be verified by the study team, as part of the process of evaluating the efficacy of the two different types of auditor (audits of the auditors). The auditor will be told this, but will not receive any other form of incentive.

<u>SIR Intervention (80 schools)</u>.

The SIR intervention will be identical to the ATE intervention except that each inspector will collect information on the presence of the head teacher, *all* teachers, and *all* students, together with the pedagogy of two randomly-selected teachers.

Scoring rubric. These four performance measures will be scored to produce an overall SIR score. In each case, scoring is based on *absolute* performance against a pre-specified objective rubric.

- Head teacher attendance: present or excused absent = 8 points, unexcused absent = 0 points.
- Teacher attendance: sliding scale from 0 to 8 points for share of teachers present (max for >90% present).
- Student attendance: sliding scale from 0 to 8 points for share of students present (max for >90% present).
- Teacher pedagogy: sliding scale from 0 to 8 points for average share of students engaged in active teaching activities (max for >88% engaged) across the two classroom observations.

The detailed protocol for all four dimensions of this scoring rubric will be communicated to the inspector during training, and to the head teacher at the start of the first surprise visit, via an information sheet.

Career incentives. The scores across these four performance measures will be aggregated into a single, cardinal score for each head teacher; specifically, the overall score will be the unweighted sum of 12 measures (head teacher attendance x3 inspections, teacher attendance x3, student attendance x3, teacher pedagogy x3), each scored on a 0-8-point scale.

As in the ATE intervention, these cardinal scores will feed into the ordinal ranking exercise within circles. At the end of the school year, the schools within each circle x SIR treatment pair will be ranked on the basis of their *cardinal* SIR score. In the event of a tie, the inspector (who, by construction, visited all schools in the comparison set) will be asked to consider all four dimensions of performance and break the tie in favor of one school or the other. The head teacher of the top-ranked school will have his/her promotion fast-tracked by one-year, while the head teacher of the bottom-ranked school will have his/her promotion delayed by one-year. The two head teachers of the two middle-ranked schools will experience no change.

Again, in an attempt to prevent dysfunctional behavior (collusion and/or demotivation) among head teachers, the precise details of the promotions process will *not* be communicated to head teachers (only a basic picture of how promotions will be tied to relative performance)..

Audits. Scores awarded as part of the SIR will also be audited, following an identical protocol to the ATE intervention.

‘Business-as-usual’ control (80 schools)

The same district inspector will be visiting schools under different treatments. This creates the possibility of spillovers from the ATE and SIR interventions into controls schools. To mitigate this, at the *start* of inspector training (before any mention of the ATE and SIR interventions) inspectors will take part in facilitated group discussions about the protocols *currently used* for teacher ‘annual confidential reports’, and for school inspections, in their district. On the basis of these discussions, a written ‘business-as-usual’ protocol will be established for each district-gender. During training it will be stressed to each inspector that it is critically important that he/she adheres to this ‘business-as-usual’ protocol for all schools in the district that have not been included in either the ATE or SIR interventions.

Additional experiment (80 schools)

Further, in a randomly-selected half of the ATE schools and a randomly-selected half of the SIR schools, an ‘inspirational video’ will be shown to teachers at the end of each inspection. The video interviews two teachers who have inspired their students to continue onto successful paths, and one student whose teacher inspired him onto a very successful path. The intention is to promote high aspirations among teachers and head teachers for changing the lives of their students for the better. We will examine whether efforts to stimulate ‘intrinsic’ motivations enhance the impacts of the ‘extrinsic’ promotion incentives.

Intervention Start Date

October, 2017

Intervention End Date

February, 2018

Outcomes

This project intends to provide three peer-reviewed papers as outputs. The first will measure the impact of the ATE and SIR interventions on a number of student and educator outcomes. The second will assess the accuracy of two different ways of measuring educator intrinsic motivation: the Perry public service scale (which was administered to all teachers and head teachers) as well as a lab-in-the-field experiment – the dictator game – which was played by all teachers and head teachers against the students in grade four (teachers and head teachers were given some money and secretly decided how much to keep for themselves and how much to donate to the students for school supplies). The third will compare the accuracy and cost-effectiveness of the two types of auditors (secondary school teachers from the same district vs. district education officials from non-study districts).

Paper 1 – Impact of ATE and SIR:
Endline test scores
Student attendance on the date of follow-up testing
Grade 4 Dropout Rates between baseline and endline
Continuation from grade 4 to grade 5 – conditional on funding
Educator attendance on the date of the follow-up testing
Teaching practices, reported by children – whether teacher provides help outside the classroom, whether teacher provides feedback to their parents, and whether the teacher splits students into groups for activities in class

Paper 2 – using games vs Perry public service scale to measure motivation, focusing on the inspirational video treatment:
Dictator game – continuous variable of the amount of money “donated” to schools
Perry public service scale
Endline test scores
Student attendance on the date of follow-up testing
Grade 4 Dropout Rates
Continuation from grade 4 to grade 5 – conditional on funding
Educator attendance on the date of the follow-up testing

Paper 3 – Which type of auditor is better:
Difference between the inspector’s score and the auditor’s score (number of points) for a full visit – for each of ATE and SIR
Difference between the inspector’s score and the auditor’s score (number of points) for each aspect measured (teacher attendance, head teacher attendance, student attendance, pedagogy, and student learning) – for each of ATE and SIR
Cost of one district official audit
Cost of one secondary school teacher audit

Experimental Design

At the start of the Pakistani school year, in September 2016, we will visit 240 government primary schools in rural areas of Charsadda, Mardan, and Nowshera districts. In all schools, we will administer a mathematics examination to grade 4 students that will cover material in the Pakistani math curriculum from grades 1-3. We will interview each grade 4 student, the grade 4 math teacher, and the head teacher in each school.

Towards the end of the school year, in February 2017, we will re-visit the 240 schools visited at baseline and administer another mathematics examination to the same students. The exam will cover material from the Pakistani curriculum from grades 1-4. No questions will be repeated. We will also administer the student, grade 4 math teacher, and head teacher questionnaires again, possibly with additional items.

In 80 of the sample schools, the grade 4 math teacher will receive a letter from the Education Department during the first inspection visit from the inspector outlining the new ATE that will substitute for the PER for this teacher. It will describe the components that will be measured, the frequency, who will measure them, and how the scores will affect their promotions. In a further 80 sample schools, the head teacher will receive a similar letter describing the new SIR.

In the remaining 80 control schools, school inspections and teacher performance evaluations will carry on as usual.

The training of inspectors and auditors was carried out in two sessions, the first from October 12-14, 2016, and the second from October 17-19 2016 (to allow inspectors to attend on the most convenient day for them). First, the training participants were given a questionnaire to fill out on their own (on demographics and the Perry public service scale), and they also played the lab-in-the-field dictator game. Next, groups were formed by district and gender to discuss and write down how school inspections are currently performed. These documents became the ‘control group protocol’. Next, the group was explained the background and purpose of the intervention and was trained on the protocol and questionnaire (done on tablets) of how to perform the ATE and SIR. They also practiced these.

Randomization Method

Sampling of 240 schools. Following discussions with the E&SE Department, three districts (Charsadda, Mardan, and Nowshera) were identified as locations where it would be logistically feasible to operate. Enrolment data from Pakistan’s Independent Monitoring Unit identified 1,547 schools satisfying the following criteria:

- Rural, public, primary school
- Grade 4 enrolment between 10-50 inclusive (based on Fall 2016 enrolment)
- Grade 4 teacher neither up for promotion nor transfer within next two years, not the head teacher, and not having the Pakistan Reading Program (PRP) in place.

Dropping schools used in the same district as a pilot, leaves 870 schools for sampling—629 males and 241 female. These schools are organized into *circles*, which are administrative units that operate much like a school district; within each circle, there is a set of boys' schools and a set of girls' schools, each with separate E&SE department officials overseeing their administration. Accordingly, there are two circle-genders within each circle—one boys' circle-gender and one girls' circle-gender.

A total of 20 circle-gender pairs will be drawn from the 48 circle-gender pairs available. To be drawn, a circle-pair must satisfy the requirement that there are at least 18 schools (to allow for replacements if needed). Given the relative scarcity of girls' schools, this requirement will result in 5 girls' circles and 15 boys' circles. Within each circle-gender pair, 12 schools will then be drawn for inclusion in the study. 12 schools x 20 clusters gives the 240 schools in the study.

Assignment of schools to treatment (ATE, SIR, control), and assignment of auditors to inspectors (official from non-study district or secondary school teacher) will be done in office, by computer by the PI team. Randomization was stratified by circles: 4 schools in each circle were assigned to the ATE treatment, the SIR treatment, or control. The randomization was carried out 1,000 times, and the balance of each randomization tested along 18 student and teacher variables collected from baseline. The randomization iteration with the minimum-maximum t statistic was the randomization iteration that was chosen.

Other assignments required as part of the intervention (e.g. choice of classroom to visit for teacher pedagogy observations under the SIR treatment) will be done in the field, automatically by the tablet on which the inspection is being undertaken using data entered by the inspector upon arrival at the school.

Randomization Unit
School

Was the treatment clustered?
yes

Sample Size: planned number of clusters
20 circles

Sample size: planned number of observations
240 head teachers, 240 grade 4 math teachers, 6,000 grade 4 students

Sample size (or number of clusters) by treatment arms
Control: 80 schools, head teachers, and grade 4 math teachers. 2,000 students.
ATE Treatment: 80 schools, head teachers, and grade 4 math teachers. 2,000 students.
SIR Treatment: 80 schools, head teachers, and grade 4 math teachers. 2,000 students.

Institutional Review Boards
IRB Name: International Food Policy Research Institute (IFPRI) Internal Review Board
IRB Approval Date: **08/23/2017**
IRB Approval Number: **17-08-23**

Analysis Plan

Our primary outcomes relate to students. The first is student learning, as measured by the difference in performance of students in Grade 4 math classes at endline compared to baseline on written, independently administered math tests. The second is student drop out, measured as a student having been enrolled in the class at baseline but not at endline. The third is student attendance, as measured on the day (and day before) the baseline and endline surveys. These will be constructed as described below.

Student learning. Grade 4 math students in all 240 schools will sit a written math test at baseline in September 2017 and again at endline in March 2018. These test scores will be used to obtain estimates of student learning (using item response theory), which will be used to test for treatment impacts in a (now standard) ANCOVA student-level specification.

Student drop out. We define drop outs as students who enroll at the start of the school year but either stop attending during the year and are withdrawn from the register or fail to enroll again the following school year (the latter outcome being conditional on funding availability for a second follow-up).

Student attendance. We measure student attendance on the day of the baseline survey as well as on the previous day. We again collect same day and previous day attendance at endline. These provide us with two measures of attendance.

Our secondary outcomes relate to teacher and headteacher effort, namely head teacher attendance and grade 4 math teacher attendance as well as teaching practices as reported by students.

Head teacher attendance. Head teacher attendance is recorded on the day of the (unannounced) baseline and endline surveys.

Teacher attendance: During surveys, enumerators will ask the head teacher for access to the official teacher register. From this they will record whether the teacher was present, absent but sanctioned, or absent but unsanctioned on each of the previous 5 school days. If there is no record, the enumerator will ask the head teacher to provide a response. Our measure of grade 4 math teacher attendance will average over these 5 days. (This is a different measure of attendance from those collected during ‘surprise visits’ in ATE and SIR schools as part of the intervention.)

Teaching practices. These will be reported by children – indicator variables for whether teacher provides help outside the classroom, whether teacher provides feedback to their parents, and whether the teacher splits students into groups for activities in class (averaged over students in the teacher’s class since the regression will be at the teacher level).

Head teacher helps with grade 4 class. To determine whether the interventions generated more of a focus on grade 4 (that was the class that was tested), we will also examine the effects of the interventions on a dummy variable for whether the head teacher helped with teaching the grade 4 class.

We will examine treatment heterogeneity along the following dimensions:

1. Baseline student achievement levels
2. Student gender
3. Educator gender
4. School resources
5. Child household resources

The data on each of the outcomes listed above should contain roughly 6,000 observations from 240 different schools. We should have significant power to detect treatment heterogeneity. Most of our key results will be treatment-control comparisons among sets of cell means, but we cannot know exactly how we will define these cells before we see the distributions of baseline outcomes.

Our empirical specification will be as follows:

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

where Y_{it} is an outcome for unit *i*, which will either be 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. We include Y_{it-1} as a control for the baseline value of the outcome, given the low level of autocorrelation these outcomes tend to have (McKenzie, 2012). School district fixed effects, γ_c, are included as randomization was stratified by school district. Standard errors are clustered at the school level; the level at which treatment was assigned (Abadie et al, 2017).