Study Title: Evaluating the Design and Impact of the Secondary School Teacher Training Initiative under the Government of Nepal's School Sector Development Program

Pre-Analysis Plan

April 2019

Julie Schaffner (Tufts University)
Paul Glewwe (University of Minnesota)
Uttam Sharma (Independent Consultant)

I. Introduction

This document presents the pre-analysis plan for a Randomized Control Trial (RCT)-based quantitative study of the impacts on teaching and learning outcomes in 9th and 10th grade math and science of teacher training programs rolled out by the Government of Nepal under the School Sector Development Program (SSDP). This quantitative study is part of a larger mixed methods research project. Schaffner, et al. (2018) describes the study design and the findings from baseline data collection.

II. Motivation and Study Description

Nepal has made great strides in raising school enrolment in recent years, but average student learning in Nepal's public schools remains low. Recognizing that development success requires the country's children and youth to acquire valuable math, science and language skills, the Government of Nepal has prioritized efforts to improve school quality over the seven years of the School Sector Development Program (SSDP), 2016-2023. Observing that many teachers (especially at the secondary level) remain weak in subject content, and that many teachers continue to use highly teacher-centered pedagogical practices, such as lecturing from the blackboard with little student engagement, policymakers hope to improve teacher effectiveness by providing them with training to improve their knowledge of subject content and to encourage them to use more engaging teaching practices.

This RCT-based study is designed to evaluate the impacts on the subject knowledge and pedagogical practices of teachers of 9th and 10th grade math and science, and on student learning, of SSDP teacher training (TT) and to examine its theory of change. The goal of this study is to provide guidance for future policy decisions regarding the scaling up or re-design of these policies. At the request of our government collaborators, we focused the study on government schools that include at least grades 1 through 10, which are considered models for what most schools will soon be in Nepal.[1]

The interventions. The main teacher training (TT) intervention requires all 9th and 10th grade teachers in government schools to attend government-run in-service teacher training modules that are intended to raise their subject knowledge and to motivate and equip them to use practical, demonstration-based teaching methods rather than more traditional teaching methods. The training includes a 10-day session at an Education Training Center (ETC), and then completion by participating teachers of the equivalent of five days of "self-study project work," which includes independent lesson plan development and other classroom research and practice activities, on which they must submit a report within 45 days of completing their training at the ETC.

[1] Schools that include at least grades 1 through 10 constitute approximately 20% of all schools in Nepal, many of which include only primary grades (Government of Nepal, 2016). According to EMIS data, however, approximately 97 percent of grade 9 and 10 government schools students are found in schools that include at least grades 1 through 10.

ETC training sessions take place during the regular school year. Teachers are provided with per diems for their stays at the ETCs.

The TT intervention in study schools differs slightly from the broader intervention to be rolled out throughout Nepal over the next several years in several respects. First, rather than waiting for teachers and schools to request trainings, the ETCs specifically invited teachers in treatment schools, and were asked not to invite teachers in control schools or other schools within the same small geographic areas (associated with the local Village Development Committees) as the control schools. Second, while in the broader roll-out priority will be given to inviting teachers with permanent positions who had not received training under the previous education plan (the School Sector Reform Program or SSRP), ETCs were requested to invite all teachers of grade 9 and 10 math and science, regardless of contract type or previous training experience.[2] Third, while in principle the full SSDP training will include two modules, each including 10 days of training at an ETC and five days of self-study project work, in practice only the first module has so far been rolled out, and only this first module is included in this evaluation.

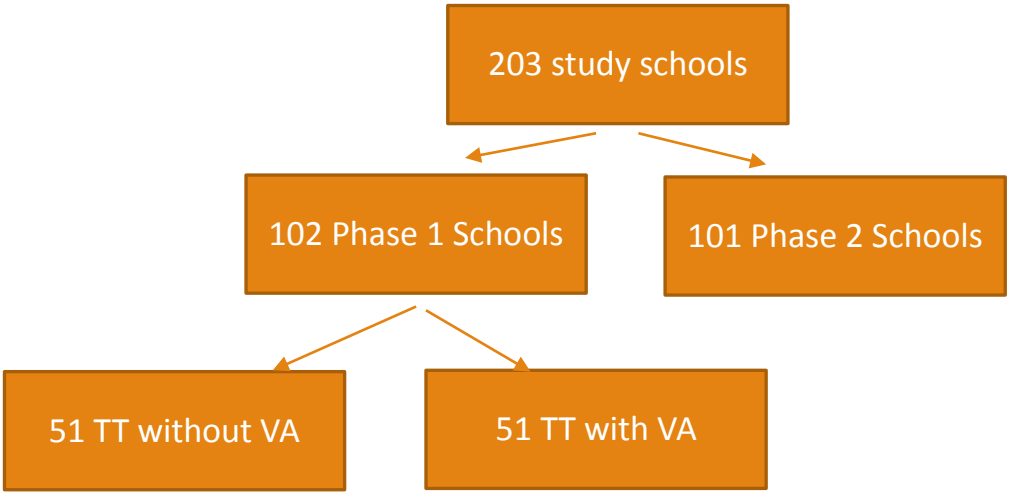
The supplementary Video Assignment (VA) treatment requires each trained teacher to submit (before he or she can receive full credit for the training) a video of himself or herself (during an entire class session) implementing one of the new lesson plans that he or she is expected to create as part of the self-study project. The aim of adding the VA is to increase teachers' motivation for investing serious effort in the self-study project activities, which may be important for translating what teachers learn at the ETC into new and improved classroom practices. In what follows, "TT treatment" will refer to the provision of the basic teacher training intervention without the video assignment, while "TTVA treatment" will refer to the provision of the teacher training intervention with the video assignment.

Sample size and study arms. Power calculations suggested the need to include approximately 100 treatment and 100 control schools to estimate the impact of the TT treatment on student test scores with adequate power.[3] Budgetary limitations prevented the addition of another 100 schools for the TTVA treatment. The research team chose, therefore, to divide the TT study arm into two sub-arms, with one receiving only the TT treatment, while the other receives the TTVA treatment. The primary randomization, therefore, divides the study schools into two groups of equal size: 1. "Phase I" schools, which were to receive the SSDP teacher training in late 2017; and 2. "Phase II" schools, which were to receive the SSDP teacher training only after May of 2019, and which serve as the control group during the period of study. To minimize the potential for spillover effects of training on Phase II study schools, other schools in the same small geographic areas that contain the Phase II schools will also receive the relevant training only after May of 2019. The secondary randomization divides Phase I schools into two groups: 1. TTVA schools, in which each teacher must submit a video of himself or herself implementing in his or her classroom a new lesson plan developed as part of the SSDP training to receive full credit for the training; and 2. TT schools, in which no video is required to receive full credit for the training. Figure 1 illustrates this basic study design.

[2] Early SSDP documents suggested that the SSDP teacher trainings would be significantly different from previous trainings, and thus that they would have the potential to improve teaching and learning outcomes even for teachers who had received SSRP training. On-going process evaluation work suggests that the differences between SSDP and SSRP training may be smaller than initially anticipated.

[3] Power calculations were done to choose a sample of schools sufficiently large so that the evaluation would have at least an 80 percent chance of detecting (at the 95 percent significance level using a two-tailed test) an impact of the TT intervention on average student test scores of at least 7 percentage points (about 0.3 standard deviations of the distribution of test scores). Estimates based on existing databases of Nepalese student standardized test scores suggested that a sample of at least 200 schools would be required to achieve this level of power.

Figure 1: Randomization of the Study Schools

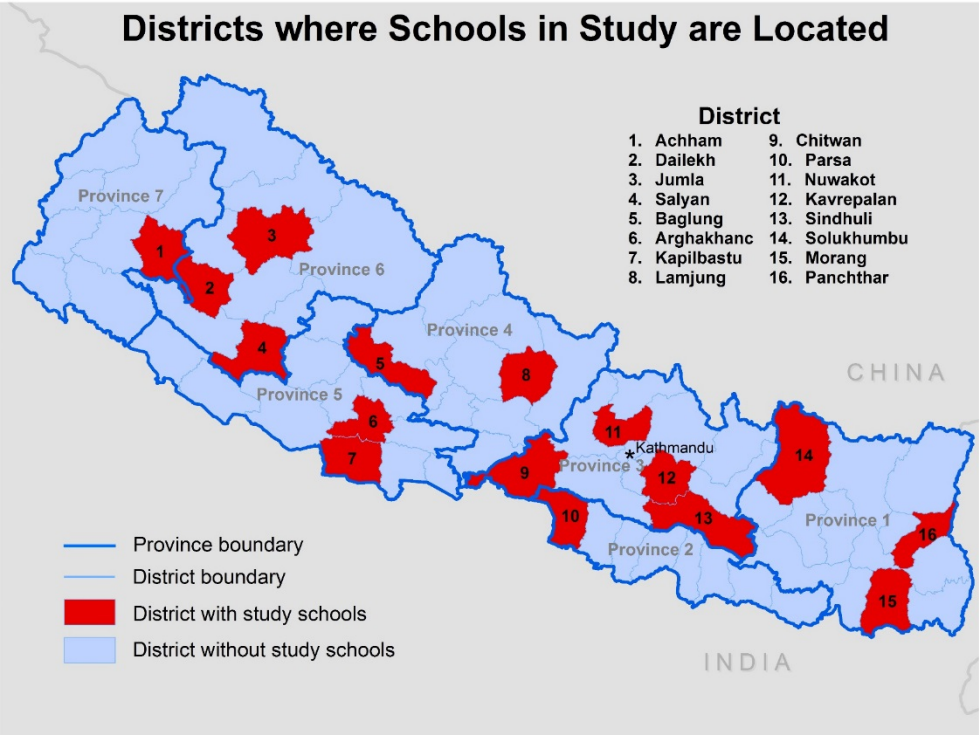


In light of experience with the administration of academic achievement tests at baseline, we chose also to randomize schools at endline, in cross-cutting fashion, to different modes of test administration, related to the timing of the assent process and to the order of the math and science assessments. Within district and study arm we randomly allocated one-third of schools to have the informed assent process for students administered before the assessments (as is standard, and as was done at baseline, but which may have made the low stakes nature of the assessments especially salient to students), while allocating the other two-thirds of schools to have the assent process administered immediately after the assessments (and before students submitted their assessment papers). Our interest in delaying the assent process until after the assessments arose out of the observation of low scores, and enumerator reports of poor assessment-taking discipline, at baseline. We retained the baseline assent process in one-third of the schools, however, so that we can evaluate the extent to which the change in assent process contributes to any improvement in test performance.

Also within district and study arm, students in half the schools were assigned to take the math assessment first, while the other half were assigned to take science assessment first. At baseline students in all schools took the math assessment first. Because test-taking fatigue may reduce performance on the second test relative to the first, randomizing which assessment is given first improves the ability to compare students' performance across subjects.

Stratification by district. To achieve a sample representative of most of Nepal, while containing costs, the research team chose to stratify the sample first by district. To reduce data collection costs further, the team (in consultation with its partners in the Nepalese government) eliminated from consideration 10 of the most remote or otherwise difficult districts. From the remaining 65 districts (which contain 94.3% of Nepal's schools), the team randomly selected a representative set of 16 districts, and then sampled schools only within those districts. The details of how this was done are described in Schaffner et al. (2018). The 16 selected districts are shown in Figure 2.

Figure 2: Sixteen Selected Districts



Stratification by previous training experience. At the request of the team's government partners, and to increase the study's power to identify impact on the teachers for whom the TT training was most likely to have impact, the team sorted schools within districts into two strata: "priority" and "non-priority" schools, and over-sampled the former. Schools were identified as priority schools if there was no evidence (in hard copy records made available by the National Center for Educational Development) of any permanent teacher, or any teacher for which contract type was unknown, having completed all three modules of the SSRP training. The specifics of this rule were dictated largely by the idiosyncrasies of the only existing records identifying which schools and teachers had received SSRP training. Further details on the process can be found in Schaffner et al. (2018). The team chose to select two-thirds of the sample within a district from the priority stratum, while the other one-third would come from the non-priority stratum.

Sample structure. The study sample was designed to: 1. Be large enough (according to power calculations) to yield sufficiently precise impact estimates; 2. Be approximately representative of all schools in Nepal that include at least grades 1 through 10; 3. Ensure that most of the sampled schools have teachers who had not completed SSRP training; and 4; Be easily divisible into halves, thirds, and quarters within districts, so that (a) half of the schools within each district and stratum could be allocated to implement the SSDP program during the study period while the other half would implement the SSDP only after the study period; (b) two thirds of the sample could be drawn from a priority stratum and one-third from a non-priority stratum (these two strata were described above); and (c) one half of the schools receiving SSDP training (i.e. one quarter of the entire sample within a district) could be assigned to also receive the TTVA treatment. The following paragraphs summarize the sample selection process. For more details, see Schaffner et al, (2018).

To facilitate selecting two-thirds of the sample from the priority stratum and one-third from the non-priority stratum, it was useful to select a number of districts per school that is divisible by three. To facilitate allocating one-fourth of the sample each to TT and TTVA treatments (while allocating the other half to control), it was useful to choose a number of schools per district that is also divisible by four. The team chose, therefore, to select 12 (a number divisible by both 3 and 4) schools per district in most districts. At our government partners' request, we doubled the number of schools in one of the larger districts (Morang). We thus aimed for a total sample size of (12*15 + 24 =) 204 schools.[4] In the end, the sample included only 203 schools, because one district, Solukhumbu, had only 3 non-priority schools, rendering it impossible to choose 4 non-priority schools there.

Selecting schools within districts and priority/non-priority strata. Concerned about potential spillovers of impact from Phase I (treated) schools to untreated schools near to them geographically, the research team chose to select priority and non-priority schools within districts in a way that would reduce the probability of any two sample schools being near to each other geographically. Rather than take simple random samples from among lists of priority and non-priority schools in a given district, therefore, the team first grouped schools by the Village Development Committee (VDC) territories to which they belonged. VDCs are administrative sub-units within districts. In the 16 districts from which the schools were drawn, there were 1,251 "eligible" schools (i.e. schools that had at least grades 1 through 10 and had not apparently received SSDP training) spread over 751 VDCs, so the average VDC had 1.67 eligible schools. The general process to reduce the probability that two sampled schools would be very close to each other was to draw a sample of VDCs, and then within each VDC randomly draw only one school.

Assignment of Schools to Study Arms. Within district and priority/non-priority strata, one-fourth of schools were allocated to Phase 1 TT treatment (SSDP training), one-fourth to Phase 1 TTVA treatment (SSDP training with Video Assignment), and one half to "Phase 2" treatment (which serves as control). Of the total sample of 203 schools in 16 districts, 51 were randomly assigned to receive the TT treatment without Video Assignment, 51 were randomly assigned to receive the TT treatment with Video Assignment, and 101 were randomly assigned to Phase II.

Sampling weights. To produce estimates of the mean or variance of a variable, or the average of a heterogeneous effect, for the population of schools (with at least grades 1 through 10) in the 16 study districts (which in turn are representative of most of Nepal), it is necessary to employ sampling weights that adjust for differences in the number of schools per district and differences in shares of priority and non-priority schools across districts. The research team used Monte Carlo methods to calculate the appropriate weights.[5] In what follows we use the label "weight1" for these weights, which are appropriate when studying school- or student-level outcomes.

The larger mixed methods study design. While this pre-analysis plan focuses on the methods we will use when analyzing data collected as part of the RCT just described, it is important to note that the RCT is part of a larger mixed methods study involving the following components: (1) A preliminary qualitative study conducted in February-April 2017, aimed at informing the quantitative study design; (2) Qualitative research conducted in June-August 2017 aimed at refining details of the Video Assignment; (3) Baseline data collection for the RCT conducted in August-December 2017; (4) In-depth in-person interview study of intervention roll-out with teachers, trainers, and other actors in three districts, conducted in October-November 2018; (5) Telephone interview study of intervention roll-out with teachers and trainers in all study districts, conducted in January-

[4] Morang district was selected to have the "double" sample because it is the largest district in the sample; in the administrative data used to select the schools, Morang has 154 of the eligible 1,334 schools in the 16 selected districts, which is more than any other district.

[5] Monte Carlo methods were required to account for the complex structure of sampling without replacement.

February 2019; and (6) Endline data collection for the RCT, conducted in February and March of 2019, for which data are currently being digitized. The research team has sought to exploit important complementarities between quantitative and qualitative research, with the aim of not only estimating SSDP training impacts but also examining theories of change and illuminating related governance challenges, for the purpose of informing future policy making in Nepal in the area of education. Through preliminary qualitative research we refined our evaluation questions and investigated how best to measure school management quality, school management practices, and other important control variables during baseline and endline quantitative data collection. Through more broadly exploratory qualitative research, we assisted the government with refining the details of the VA treatment. Through in-depth in-person interviews with teachers, trainers, video assignment focal persons and teacher union representatives, and through qualitative telephone interviews with broader samples of teachers and trainers, we have identified points of strength and weakness in the implementation of the interventions and in their theories of change. The qualitative components have played important roles in shaping the design of the quantitative research, and will also play important roles in helping us interpret the quantitative results.

Data collected at baseline. At baseline data were collected over the period August 2017 to January 2019, using the following instruments:

- A head teacher questionnaire administered by an enumerator using a tablet. The questionnaire includes questions about the school, the head teacher, the school management committee and school management practices, as well as questions about teachers and teaching practices in 9th and 10th grade math and science.
- A questionnaire for teachers of 9th and 10th grade math and science administered by an enumerator using a tablet. The questionnaire includes questions on teacher characteristics and school management practices.
- A student questionnaire that students are asked to fill out themselves on paper copies, for students in grades 8 and 9 (who will be in grades 9 and 10 at endline). The questionnaire includes questions on student socioeconomic characteristics and the teaching practices employed in their math and sciences classes.
- Student assessments in math and science. Each student in grades 8 and 9 took two one-hour assessments, one in math and one in science. The assessments are standard math and science assessments that were developed for this study by Nepalese professors of education at Tribhuvan University. The assessments are similar to the standardized tests administered by the government, including both multiple choice and open response questions.
- Measures of classroom teaching practices and student engagement based on the Stallings method of classroom observation (World Bank, 2017*)*.[6]

Data collected at endline. At endline, conducted during February and March of 2018, data were collected using the following instruments:[7]

- A head teacher questionnaire administered by an enumerator using a tablet. The content is similar to the content at baseline.

[6] See http://documents.worldbank.org/curated/en/733701505747664220/pdf/119754-REVISED-PUBLIC-WBManualENGV.pdf for further information on the Stallings method.

[7] Endline data were being collected, digitized and cleaned during the refinement of this pre-analysis plan. The cleaning will be done using datasets stripped of study arm indicators.

- A questionnaire for teachers of 9th and 10th grade math and science administered by an enumerator using a tablet. The content is similar to the content at baseline.
- A student questionnaire that students are asked to fill out themselves on paper copies. The content is similar to the content at baseline. Enumerators read each question aloud and answered student questions, to encourage accurate and complete responses.
- A School Management Committee respondent questionnaire. This questionnaire focuses on the relationship between the school and the new local governments that have recently been elected as part of the broad "federalization" of government that is taking place in Nepal.
- Student assessments in math and science. Each student in grades 9 and 10 took two one-hour assessments, one in math and one in science. The assessments for endline were drafted by U.S.-based consultants with expertise in psychometrics, who were asked to construct assessments that would be tailored to the Nepalese curricula for grade and subject, paying special attention to curriculum content emphasized during the SSDP trainings, while also including questions at lower grade levels (to assess how many students are entering grades 9 and 10 with preparation below grade level). The consultants were asked to draw on questions from the baseline assessments (allowing linking to those assessments) as well as question banks from international assessments (allowing incorporation of high quality items that had already been refined through intensive pre-testing). They drafted two very similar assessments (called "Version A" and "Version B") for each subject and grade, so that the risk of students cheating by copying could be reduced (by making it possible to have students take alternating exams within a row in rooms with large numbers of students) and subject content covered could be increased. The two assessments for a given subject and grade contain linking items. The drafts were then reviewed and amended by local consultants in Nepal, to guarantee their relevance to the Nepalese curriculum and testing style. After pre-testing, 6 questions with the lowest correct response rates were dropped from each assessment to produce the final assessments with 35 items each. These assessments include only multiple choice questions. Unfortunately, printer errors resulted in the reproduction of assessments with incorrect pages (one incorrect page each) for one version each of the grade 9 and 10 math assessments. These assessments with errors were distributed in a small number of schools before this problem was detected by the survey firm. These assessments will have to be treated as additional versions of the assessments. Fortunately, they contain many linking items with the correctly formulated assessments.
- Evaluations of selected student assessment items, to be filled out by teachers of grade 9 and 10 math and science, on paper copies. Wishing to assess teacher subject knowledge without explicitly asking teachers to take assessments, the team presented teachers with subsets of the student assessment items and asked them to rate their clarity, provide the answers that they thought the item writer intended, estimate the fraction of their students who would get the item correct, and rate how well tailored the item is to the Nepalese curriculum for subject and grade. Their answers (to the question about the intended correct answer) allow assessment of their subject knowledge.
- Teacher attendance data from school log books. Enumerators were asked to locate all current grade 9 and 10 math and science teachers in the school's log book and record whether each teacher is marked as present on the current date and the previous several days. While these records are subject to manipulation by teachers, especially after they return from absences, they may be revealing about recent and current absence rates.
- Teacher attendance data from teacher questionnaire administration. Enumerators were asked to identify all current teachers of grade 9 and 10 math and science and to fill out at least a first question (regarding presence or absence) on the first day of a visit to a school (whether the teacher was there or not). The hope is to obtain another measure of attendance rates. In most cases the schools were alerted to the visits, however, so these attendance rates may be inflated relative to a typical school day.

- Student ethnicity data. While at baseline we inferred students' ethnicity from their last names, at endline the enumerator teams consulted school records or school personnel to obtain ethnicity information for students present at baseline or endline. (The team did not ask students to report their own ethnicity, because we did not want to make students' ethnicity salient to them while they were taking the assessments.)
- For students who were present at baseline, their scores on the previous year's end-of-year exams. While end-of-year exams are not standardized across schools, and thus are unlikely to be useful for estimating impact, they may be useful for studying differences in performance within schools between students who did and did not attrit from the sample.

Data collection at endline was designed to allow matching of the data for students present at endline to their own baseline data (for those who were present at baseline) and to their endline teacher's data. The matching to teacher is done by requesting from students and teachers the name or number of their "section" for a given grade and subject.

III. Research Questions

To evaluate the impact of the SSDP trainings on the outcomes of ultimate interest, we address the question:

1. *What is the average effect on math and science achievement scores, for students in 9th and 10th grade, of inviting all teachers of grade 9 and 10 math and science at their schools for participation in the TT and TTVA treatments?*

The primary outcome measures for estimating these Intention to Treat (ITT) impacts are:

- Overall achievement indices estimated using Item Response Theory (IRT) methods, as described below, linking data on both versions ("A" or "B") of the relevant endline assessment. This will be done separately by subject (math and science) and grade (9 and 10).
- Achievement indices on the subset of questions on content that was to be emphasized during SSDP trainings, again using IRT methods, linking data on both endline versions of the relevant assessments. This will be done separately by subject and grade.

To assess robustness, we will also examine impacts on:

- Raw total scores on the relevant assessments
- Scores estimated using similar IRT methods, but linking also to baseline assessments.

As discussed below in the methods section, the ITT estimates of impact on student test scores will be done both for:

- the full endline sample, without controlling for baseline scores, and
- the "panel sample" (of students present at both baseline and endline), controlling for baseline scores.

Because ITT impacts may be diminished by the failure of invited teachers to take up training, and by the transition of trained teachers out of treatment schools and possibly into control schools (in a relatively high turnover environment), we will also examine Question 2:

2. *What is the average effect on math and science achievement scores, for students in 9th and 10th grade, of their teacher's receipt of TT and TTVA treatments?*

The primary outcome measures for estimating these Local Average Treatment Effects (LATE) are the same as the outcomes for Question 1. As explained in detail below in the methods section, we will estimate these effects by linking students to indicators of whether their teachers in fact received the TT or TTVA treatment, and instrumenting these indicators with the school's treatment assignment indicators. Again, we will perform this estimation for the entire endline sample without controlling for baseline scores and also for the panel sample controlling for baseline scores.

Aiming to provide policymakers not only with estimates of the sizes of impact on student learning outcomes, but also with insights into why the estimates are large or small, the research team also seeks to address the following two questions about intermediate outcomes:

3. *What are the average effects on math and science teachers' attendance, subject knowledge, and pedagogical practices, and on student attendance, of inviting a school's teachers to participate in the TT and TTVA treatments?*

4. *What are the average effects on math and science teachers' attendance and pedagogical practices of their participation in the TT and TTVA treatments?*

Question 3 considers ITT estimates, while Question 4 considers LATE estimates, for intermediate outcomes that can be linked to specific teachers. Examining the impacts on these intermediate outcomes is useful, because impacts of teacher training could be weakened by problems at various point along the logical chain linking training to learning. Study of intermediate outcomes might reveal, for example, that training leads to significant impacts on teacher subject knowledge and teaching practices, but nonetheless has little impact on student learning. Such a pattern would point to the importance of understanding better the barriers that prevent students from learning, and whether and how training content might be better targeted toward removing or at least reducing those barriers. Alternatively, study of intermediate outcomes might reveal instead that training leads to significant impacts on subject knowledge but no impact on teaching practices, suggesting the need to understand better how teachers are, or could be, held accountable for implementing new practices in their classrooms, and what obstacles they face to doing so.

The intermediate outcomes to be examined include:

- School-level averages of teachers' scores on subject knowledge assessments (implicit in the teacher evaluations of the student assessment items), estimated using IRT methods (with raw scores as robustness check), separately for math and science.
- Teacher attendance measures:
    - Teacher attendance on the first day of the school visit, and over the previous several days, as recorded in school logbooks
    - Teacher attendance on the first day of visit, as recorded in the first question on the teacher questionnaire
    - Median of students' report regarding how frequently the teacher is absent (5-point scale)
- Student attendance:
    - Student attendance rate on the most recent full day of classes before today, as reported by head teacher, separately for grade 9 and grade 10 students
- From Head Teacher reports on individual teachers:
    - The teacher's command over math or science subject matter (5-point scale)
    - The teacher's interest is in learning ways to teach more effectively (4-point scale)

- Whether the teacher has ever created from local resources any teaching materials – such as models of 3-dimensional shapes, atoms or the solar system – for use in a grade 9 or 10 math or science course
- Frequency of teacher's use of teaching materials or other visual aids (other than the chalk board) to help explain concepts (4-point scale)
- Whether teacher has ever collected information from local residents outside the school (such as prices in local markets or interest rates offered by local bankers), or required his students to collect such information
- Frequency of teacher requiring students to work together in small groups (4-point scale)
- Frequency of teacher requiring students to work on longer-term project, for which they must gather information and make practical application (4-point scale

- From teacher questionnaires:
  - Self report of minutes spent preparing per class during the most recent full week of classes (calculated from reports of total number of classes and total minutes spent preparing, whether during other class periods or outside of class time).
  - Self report of whether teacher uses a written lesson plan as a primary guide while conducting a class
  - Self report of how often teacher has students work together in small groups (5-point scale)
  - Self report of how often teacher uses examples or homework problems involving local information (7-point scale)
  - Opinion regarding importance of using examples involving local information (3-point scale)
  - Self report of how often teacher has required students to collect local information, whether by interviewing family or community members, observing family or community activities, or taking measurements (for example, of weather conditions)? (7-point scale)
- From student questionnaire, median responses, separately for math and science teachers
  - How often teacher gives homework (5-point scale)
  - How often teacher checks student's homework (5-point scale)
  - How often teacher returns student's homework with corrections (5-point scale)
  - How often teacher uses class time for asking questions of any students or holding discussions or interactions about math/science concepts with any students (5-point scale)
  - How often teacher uses class time to ask YOU (the student) a question or engaging YOU in a discussion of interaction about math/science concepts (5-point scale)
  - How often teacher requires the student to work together in small groups (5-point scale)
  - How often teacher uses demonstrations involving physical objects made from local materials or other visual aids to help students understand math/science concepts (5-point scale)
  - How often teacher uses demonstrations involving diagrams, pictures or information from the internet (5-point scale)
  - How often teacher gives quizzes to students (5-point scale)

Recognizing that schools and teachers face diverse challenges that may prevent even well-executed training programs from having strong impacts, and that new teaching practices may have differential effects on students with different aptitude and preparation, the research team furthermore plans to examine impact heterogeneity. The study therefore addresses the questions:

5. *How do the ITT impacts of the TT and TTVA treatments on student learning (Question 1) and teaching practices (Question 3) in 9th and 10th grade math and science differ across: a) teachers who had and had not received SSRP training for secondary math or science teachers prior to baseline; b) schools with higher and lower school management quality scores; c) teachers of different contract types and experience levels;*

and d) students of different gender, caste/ethnicity, baseline test scores and quantile placements in the unconditional endline test score distribution?

6. *How do the LATE impacts of the TT and TTVA treatments on student learning (Question 2) and teaching practices (Question 4) in 9th and 10th grade math or science differ along the same dimensions mentioned in Question 5?*

We will examine impact heterogeneity by including (one dimension at a time) in our ITT regression specifications interaction terms between the treatment indicators and the measures of these potentially important dimensions of heterogeneity.

The first dimension of heterogeneity to be considered is:

- The extent of SSRP training among a school's teachers at endline. For school-level regressions, this will be measured by the fraction of the school's teachers at endline reporting having had SSRP training. For student- or teacher-level regressions, this will be an indicator of whether the relevant teacher reported having received SSRP training.

Though early discussions with policymakers indicated that the training under the SSDP would be significantly different from the training that the government had provided to teachers under the previous wave of education policy, namely the School Sector Reform Program (SSRP), in practice (as revealed by process evaluation research), while the SSDP curriculum requirements distributed to the ETCs were somewhat different from the SSRP curriculum requirements, the trainers received no new training of trainers, leaving some uncertainty regarding how much of the content was truly new. This raises the possibility that the trainings will have greater impact on teachers who had not received SSRP training, for whom the curriculum was more likely to be entirely new. Despite sampling two-thirds of the sample schools from the "priority" stratum, in which there was no evidence (from imperfect hard copy records) of teachers having completed all three modules of SSRP training, nearly 30 percent of teachers in the sample at baseline reported having received SSRP training (with comparable rates in both the "priority" and "non-priority" strata). Some teachers also reported having other government or NGO math or science training. While we don't know the exact nature of those "other" trainings, we do know that reports of these other math and science trainings are more common among older teachers with more years of experience, thus it seems likely that many of these trainings took place further in the past than the period of the SSRP. Historical complaints about government teacher training programs suggest that the other government math and science trainings were likely to have been much less practically oriented than the SSRP and SSDP trainings. Thus the main concern regarding SSDP impact heterogeneity has to do with previous SSRP training. Focus on the SSDP impact among teachers who had not had SSRP training may be important for detecting evidence of SSDP training impacts.

The second dimension of impact heterogeneity we consider involves:

- The school's "school management quality" score, as estimated using methods described in Appendix A.

The motivation for examining heterogeneity along this dimension arises out of widespread concern that teachers lack motivation or accountability for applying in their classrooms the new techniques that they learn about during trainings, and the potential for good school management to provide teachers with more of the relevant motivation and accountability. (See Appendix A for more discussion of this.) The driving force behind the "school management" input of interest to us could be the head teacher, the School Management Committee, or other leaders, but the management activities themselves tend to be carried out by Head Teachers, who visit classrooms, provide feedback, convene meetings and provide leadership in other ways.

Other dimensions of heterogeneity to be considered include:

- Teacher employment status. For school-level estimation this will be the fraction of a school's grade 9 and 10 math or science teachers who are permanent teachers. For student- or teacher-level estimation, this will be an indicator of whether the relevant teacher has a permanent position. This is of interest because permanent teachers are government employees with the equivalent of tenure. Teachers are hired under many other contractual arrangements and funding streams, which share in common a more tenuous job security, which gives their employers greater potential to condition continued employment on good performance. Teachers under non-permanent arrangements may, therefore, have greater motivation to apply new practices acquired through SSDP training.
- Teachers' years of teaching experience. For school-level estimation this will be the school-level share of grade 9 and 10 math and science teachers who have more than five years of experience. For student- or teacher-level estimation this will be an indicator of whether the relevant teacher has more than five years of teaching experience. On the one hand, teachers with less experience may have more to gain from practical training and from interaction with other teachers at training sessions. On the other hand, teachers with less experience have probably been trained more recently, and it is possible that the teachers trained further in the past (when pre-service training programs were of lower quality) are the ones who have the most to gain from being expose to new teaching ideas at the trainings.

The following dimensions of heterogeneity are relevant only for student test scores regressions.

- Student preparation and ability. We suspect that SSDP training will have less impact on students who enter grade 9 or 10 with below-grade-level knowledge or aptitude, because traditional teaching practices tend not to accommodate students who enter below grade level, and because the SSDP training did not seek to rectify this. Rather, SSDP training focused on having teachers use more engaging methods to teach grade-level content. We aim to examine this in three ways.
    - In endline student-level test score regressions (for all achievement outcomes), including interactions between treatment indicators and indicators of student's tercile of the baseline test score distribution.
    - In endline student-level test score regressions (all achievement outcomes), use generalized quantile regression estimation methods to map out differences in estimated impacts for students at different levels of the unconditional distribution of endline test scores. We do this, in addition to examining interactions with baseline test scores, because we have some concerns that our baseline test scores include a great deal of noise. For this reason we are also interested in estimating different impacts across quantiles of the endline test score distribution. We recognize that these estimates must be interpreted cautiously, if it seems possible that training could have altered students' ranks within the distributions.
    - In endline student-level test score regressions (only for the narrower achievement index derived from student performance on the subset of questions most connected to SSDP training content), including interactions between treatment indicators and an indicator of student performance on the questions pertaining to earlier grade levels. This indicator will be constructed by identifying a cut-off for scores on the earlier grade questions that divides the students into groups of roughly equal size, and setting the indicator equal to 1 if the student is in the upper group. (Item maps provided by the assessment developers allow us to distinguish between lower grade questions and questions at the grade 9 or 10 levels in subject areas emphasized during SSDP trainings.)
- Student socioeconomic characteristics. Policymakers concerned with creating a more inclusive school environment express strong interest in the extent to which policies such as the SSDP teacher trainings

widen or narrow differences in learning outcomes across students in diverse socio-economic groups. We will, therefore, examine heterogeneity of impact across student characteristics, including:

- An indicator of whether the student self-identifies as male (rather than female or third gender).
- A set of indicators allowing differentiated impacts across six caste/ethnicity/religion categories of interest in Nepal: Brahmin and Chhetri, Madhes (and other castes from the plains), Dalit, Newar, Other Janajati, and Muslim.
- An indicator of whether at least one parent has attended at least lower secondary school.
- A family asset index constructed by applying IRT analysis to students' answers to five dichotomous questions regarding whether their family has various assets. (For robustness, we will also construct the first principal component of the answers about these assets.)

IV. Estimation methods

The basic regression specifications are as follows:

ITT student assessment score impacts on panel sample. The main regression equation for studying student-level outcomes on the "panel sample" of students present at both baseline and endline has the ANCOVA structure:

$$Y_{is1} = \beta_0 + \beta_1 Y_{is0} + \beta_2 TT_{s1} + \beta_3 TTVA_{s1} + A_s\beta_A + S_s\beta_S + \varepsilon_{is1}$$

where Y is a student academic achievement outcome measure, TT is a dummy variable indicating a school randomly selected for the general teacher training, TTVA is a dummy variable indicating a school randomly selected for the general teacher training plus the video assessment, A is a vector of indicators describing allocation of school to the different assessment administration procedures (whether math test is given first or science test is given first, whether assent is requested before taking tests or after taking tests), S is a vector of district by priority/non-priority stratum fixed effects, i indexes student, and s indexes school. The subscript 1 refers to endline while the subscript 0 refers to baseline. The main specification will be estimated using weighted least squares, employing weight1 (as defined above).

It is possible that the impacts of the two treatments, TT and TTVA, are very similar. Thus for all of the regressions the hypothesis that $\beta_2 = \beta_3$ will be tested. If that hypothesis cannot be rejected, then a similar regression will be estimated that combines the two treatments into a single treatment. More specifically, a "general" treatment variable can be defined as T = TT + TTVA, and that variable can be added to the above regression while the TT and TTVA variables are dropped. If the null hypothesis that $\beta_2 = \beta_3$ cannot be rejected, this specification will have more statistical power than the above regression with TT and TTVA being added as separate regressors. This general approach will also be used for each of the regression equations described below.

For robustness checking this regression will also be run:

- without weights
- without the controls for test-taking conditions in the school
- with the addition of school-, teacher-and student-level controls (The student-level controls will be indicators of whether students report their fathers as having had at least some secondary education, whether they report their mothers as having had at least some secondary education, and a simple index of family asset ownership, as defined for balance tests below. The teacher-level controls will be indicators of whether the teacher had SSRP training, whether the teacher has a permanent position and whether the teacher has more than five years of experience. The school-level control is a measure

of remoteness, the time it takes (in hours) to walk from the school to the nearest all-weather motorable road.)

ITT student assessment score impacts on full endline sample. The main regression equation for studying student-level outcomes on the full endline sample of students has the cross-section structure:

$$Y_{is1} = \beta_0 + \beta_1 TT_{s1} + \beta_2 TTVA_{s1} + A_s\beta_A + S_s\beta_S + \varepsilon_{is1}$$

where the notation is the same as above. Again, the main specification will be estimated using weighted least squares, employing weight1.

For robustness checking this regression will also be run:

- without weights
- without the controls for test-taking conditions at the school
- with the addition of school-, teacher- and student-level controls (as indicated above)

ITT school-level outcome impacts. The main regression equation for school-level outcomes, using endline data only, has the cross-section structure:

$$Y_{s1} = \beta_0 + \beta_1 TT_{s1} + \beta_2 TTVA_{s1} + S_s\beta_S + \varepsilon_{s1}$$

where Y is now a school-level outcome variable, and the rest of the notation is as above (now without student subscripts). The main specification will be estimated using weighted least squares, employing weight1.

For robustness checking this regression will also be run:

- without weights
- with the addition of school-level controls as indicated above, plus school-level aggregates of the teacher-level controls (e.g. the percentage of teachers who had SSRP training).

ITT teacher-level outcomes. The main regression equation for studying teacher-level outcomes, on the endline sample, has the cross section structure:

$$Y_{ts1} = \beta_0 + \beta_1 TT_{s1} + \beta_2 TTVA_{s1} + S_s\beta_S + \varepsilon_{ts1}$$

where Y is now a teacher-level outcome, t is the teacher subscript, and the other notation is the same as above. The main specification will be estimated using weighted least squares, employing weight1. If interview data are missing for more than 5 percent of grade 9 and 10 math and science teachers, we will adjust the weights to account for uneven non-response across schools, multiplying it by the ratio of the total number of relevant teachers in the school to the number of relevant teachers for which interview data are available.

For robustness checking this regression will also be run:

- without weights
- with teacher-level controls (whether the teacher has a permanent position, years of teaching experience, whether the teacher had received SSRP training prior to baseline)

IV/LATE estimates. The regression specifications for LATE estimation for student- and teacher-level outcomes is as above, except that the TT and TTVA indicators of school treatment assignments will be replaced by indicators of whether the teacher (for teacher-level outcomes) or the teacher of the student (for student-level outcomes) in fact participated in the TT or TTVA treatments, and will be instrumented for using school treatment assignment indicators.

Estimation of academic achievement indices. The primary measure of overall academic achievement will be derived by estimating a 2- or 3-parameter logistic IRT model (with the 3-parameter model selected only if such estimation is feasible and a likelihood ratio test rejects the nested 2-parameter logistic model with a p-value of .05 or lower), using pooled data for both versions of the endline assessments within subject and grade.

Using item maps provided by the consultants who produced the assessments, the team will identify the subset of assessment items (on each assessment) that pertain to content emphasized by the SSDP training, and will perform IRT analysis of the same structure to estimate achievement scores on SSDP-related content.

Ordinal outcomes. Many of the teacher-level outcomes are ordinal, with categories indicating Likert-scale opinions or frequencies of activities. Some are dichotomous, while others have 3 to 7 categorical options. For dichotomous outcomes, we will use probit estimation. For polychotomous outcomes, we will first collapse categories (collapsing small categories into neighboring categories closer to the "middle" score) when categories contain fewer than 5 percent of the observations, and then use ordered probit estimation. When the outcomes are ordinal student reports, they will be aggregated up to the teacher level by taking medians. When the median falls midway between two integers, the median will be rounded up, so that the resulting median also takes only integer values. This is useful because it allows us to treat the aggregated measures appropriately as ordinal rather than cardinal, and again use ordered probit estimation.

Treatment heterogeneity associated with observed variables. Treatment heterogeneity will be assessed by introducing interaction terms between the treatment indicators and the variables describing the heterogeneity of interest. If not already present in the regression, the un-interacted variables describing the relevant dimension of heterogeneity will also be added. In the case of IV/LATE specifications, the set of instruments will be expanded to include the interactions between the school treatment allocation indicators and the variables describing the relevant dimensions of heterogeneity.

Heterogeneity of impact across quantiles of the unconditional endline assessment score distribution. To look for suggestive evidence regarding differences in the size of TT and TTVA impacts across students with higher or lower quantiles of the endline assessment score distribution, we will use the generalized quantile treatment effect model proposed by Powell (2016) to map out the sizes of treatment effects by quantile of the observed outcome distribution.

Standard errors. Because treatment was assigned at the school level, we will cluster standard errors at the school level. In specifications that include the pre-estimated measure of school management quality (described in the appendix), robustness will be assessed by calculating bootstrapped standard errors.

Student attrition. Data collected at endline will allow us to determine whether students who were present at baseline but not at endline have left school, moved to another school, were held back in a lower grade, or were still in the same school and grade but absent on the day of the assessments. If feasible, data on baseline students' end-of-year exam scores (for the previous year) may also be used to compare the average achievement of baseline students who are and are not present at endline. We will use Lee bounds to bound the estimates of impact if any of the following conditions holds: a) there are statistically significant differences in attrition rates across study arms; b) there are differences of at least 5 percentage points in attrition rates across study arms; and c) there are statistically significant or economically important differences in the statistics describing the nature of attrition across treatment arms. These statistics include: a) the percentage of attritters who are reported to be no longer in school rather than having moved to another school; and b) the school-level ratio of average scores among attritters to average scores among non-attritters.)

Multiple hypothesis testing. Because impacts will be estimated for many outcomes, we must recognize the potential to obtain at least some apparently significant impacts by chance. To account for this we will report

False Discovery Rate adjusted p-values (or q-values). We will calculate these adjusted q-values for two sets of tests: the tests of no differences across the three study arms (TT, TTVA and Control) for all ITT regressions associated with Questions 1 and 3, and the tests of no differences across the three treatment statuses (Participated in TT, Participated in TTVA, Participated in neither) for all LATE regressions associated with Questions 2 and 4.

Outcomes with limited variation. Any outcomes for which 95 percent of the observations or more have the same value will be excluded from the analysis.

V. Baseline balance

Table 1 provides descriptive statistics for key variables that might influence outcomes or that describe outcomes at baseline, and examines whether the values of these variables are similar across the two treatment groups and the control group. For each of the 23 variables in the table, two tests that the means are equal were done, using regression analysis. The first tests the equality of the three means for the three groups (TT, TTVA and control) by testing whether coefficients in a regression of the variable of interest on a constant term two dummy variables, one for TT and the other for TTVA, are (jointly) insignificantly different from zero. (These regressions also include strata dummy variables.) This null hypothesis is rejected only one out of 23 times, and only at the 10% level. The second tests whether the average mean of the two treated groups combined is significantly different from the mean of the control group, again using regression analysis. None of the 23 tests is significantly different from zero. Thus we conclude that the three groups of schools are well balanced. We nonetheless note that the point estimates of mean differences are more than trivial in size in a few cases. For this reason we maintain interest in running regressions including school-, teacher- and students-level controls, for the sake of assessing robustness, as indicated above.

VI. Miscellaneous

Data access. The data and do-files used in the analysis will be published on the 3IE dataverse, and possibly in other locations, within 1 year after the completion of endline data collection. (More specifically, the research team plans to submit a final report to 3ie within 6 months after completion of endline data collection, and to publish data and do-files within 6 months after that.)

Human subjects research ethics review. This research was approved by Institutional Review Boards at Tufts University and the University of Minnesota.

Appendix A: Construction of school management quality index

This appendix describes the steps taken to develop and evaluate the measures of *school management quality* employed for studying the heterogeneity of impact (across schools with different levels of baseline management quality) of a Government of Nepal teacher training program for teachers of grade 9 and 10 math and science.

In broad brush, we treat the problem of estimating schools' levels of management quality as analogous to the problem of estimating students' levels of academic achievement. Modern psychometric best practice for estimating an index of students' levels of mathematics achievement involves: (1) developing mathematics achievement assessment items, (2) administering multi-item assessments to students, (3) evaluating the validity and reliability of the assessment items, and (4) estimating and evaluating an Item Response Theory (IRT) model relating a student's observed assessment item responses to their levels of a latent mathematics achievement measure. In a similar way, we estimate schools' level of school management quality by (1) developing a set of "school management quality assessment items," (2) gathering data on these assessment items for each school in our sample at baseline, (3) evaluating the validity and reliability of the assessment items, and (4) estimating and evaluating an IRT model that relates a school's observed assessment item values to the school's level of a latent school management quality measure.

Motivation and overview

Recent attempts to measure school management quality focus on measuring the intensity and quality of school management practices, using open-ended questions administered to head teachers (school principals). For example, working with highly educated and intensively trained enumerators administering phone interviews with head teachers, Bloom et al.'s (2015) World Management Survey (WMS) gives scores in 20 areas of management practice to over 1800 schools educating 15-year-olds across eight countries, and then use the average scores across the 20 areas as their index of school management quality.[8] Enumerators were trained to use the WMS tool using open-ended questions and conversational demeanor, in the hope of preventing respondents from realizing that they are being evaluated. Within each practice area, higher scores connote practices that are more structured in ways that reflect goal-oriented design and school-wide application, together with consistent monitoring and oversight. The authors find that the values of their management quality index vary significantly both across and within countries and that a one standard deviation increase in their index is associated with a 0.2 to 0.4 standard deviation increase in pupil outcomes within countries.

With the aim of increasing the usefulness of the WMS approach for contexts in developing countries, Lemos and Scur (2016) adapted the WMS approach with the aims of (a) achieving greater sensitivity to variation among schools in the lower tail of the management quality distribution (where most schools tend to be clustered in developing countries), (b) using interview methods that require fewer judgment calls by enumerators (thereby reducing somewhat the skill and training requirements for enumerators), (c) using face-to-face rather than telephone interviews, and (d) distinguishing scores within each of the 20 management practice areas for "implementation," "usage" and "monitoring" of the relevant processes. By requiring interviewers to give schools separate scores for implementation, usage and monitoring within each practice

[8] Their World Management Survey groups the 20 areas in which schools were scored into four practice areas as follows: *Operation* (standardization of instruction planning processes, personalization of instruction and learning, data-driven planning and pupil transitions, adopting educational best practices), *Monitoring* (continuous improvement, performance tracking, performance review, performance dialogue, consequence management), *Target Setting* (target balance, target interconnection, time horizon of targets, target stretch, clarity and comparability of targets) and *People Management* (rewarding high performers, fixing poor performers, promoting high performers, managing talent, retaining talent, creating a distinctive employee value proposition).

area, rather than requiring them to give overall scores within practice areas, the authors hoped to reduce somewhat the need for interviewers to make difficult judgment calls. In practice, though, the method still requires interviewers to use open-ended questions and make many judgment calls, and thus still requires intensive training and double scoring to maintain reliability.

The approach taken to construct an index of school management quality in this paper differs from the WMS approach, for several reasons. First, as discussed in the next section, the notion of school management quality that we wish to measure is somewhat narrower, relating more specifically to the management of teacher classroom practices. Second, for budgetary and logistical reasons, we chose to work with closed-ended questions, so that local enumerators with only Bachelors' degrees (many but not all of whom had experience in the education sector) could be well-trained to administer the questionnaires reliably. Third, given our particular interest in the management of teachers and how this might shape teachers' incentives toward teaching and adopting new teaching practices, we chose to use observation and opinion questions asked of teachers (shedding light on their experience of others' management practices and leadership), rather than using management practice questions asked of head teachers. Making use of teacher rather than head teacher observations about head teachers' managerial practices also has the advantage of reducing concerns about self-serving reporting biases, and possibly increasing the precision of measurement at the school level (through averaging responses across multiple teachers, if measurement errors are independent or only weakly correlated across teachers). Fourth, because the school management environment in Nepal's government schools was largely unstudied, we preferred to remain agnostic about which questions would ultimately be included in the index, and about the weights that would be placed on the various answers when constructing the index, and to instead use Item Response Theory model estimation to guide index construction (following best practices from the psychometric literature).

More specifically, we treat school management quality as a latent variable, the values of which must be estimated using data on an array of observed dichotomous and polychotomous "school management quality assessment items," which are relatively easy to ask and easy to answer. The items are measured as school-level medians across responding teachers of grades 9 and 10 math and science (where medians that fall between integers are rounded up to the next integer). Denoting school i's latent school management quality by θ_i, and letting Q_{ij} be the dichotomous or polychotomous response to management quality assessment item j for school i, we assume that

$$(1) \qquad \text{Prob}\,(Q_{ij} \geq k \mid \theta_i) = \exp\{a_j(\theta_i - b_{jk})\} / (1+\exp\{a_j(\theta_i - b_{jk})\}) \quad \text{for all i,j,k.}$$

where k=0,1,…,K are the values the categorical responses might take.[9] The "units" of the latent management quality variable θ_i are set by combining equation (1) with the normalizing assumption that θ_i has a standard normal distribution (so that most schools will have values of θ between -2.5 and 2.5). Item Response Theory models of this sort are similar in spirit to factor analysis models, but are more appropriate when the observed items are dichotomous and polychotomous rather than continuous.

For insight into the nature of the model and the interpretation of its parameters, it is useful to examine the special case in which all candidate test questions have dichotomous responses. In this case Q_{ij} equals 1 (0) if characteristic j is present (absent) in school i. When this is the case, the general model of equation (1) above reduces to the "two-parameter logistic model," which may be expressed as

$$(2) \qquad \text{Prob}\,(Q_{ij}=1 \mid \theta_i) = \exp\{a_j(\theta_i - b_j)\} / (1+\exp\{a_j(\theta_i - b_j)\}) \quad \text{for all i,j.}$$

[9] For simplicity, the notation treats all test questions as having the same number of possible responses, 0-K. In practice, however, the questions may have differing numbers of responses, from two (0 and 1) to a maximum (in our case 5).

Letting T be the number of assessment items, parameters $a_1,\ldots a_T$ are item-specific "discrimination parameters," and $b_1,\ldots,b_T$ are the item-specific "difficulty parameters" (as will be explained below).

This model has the appealing implication that as the value of school i's latent management quality (θ_i) rises, the probability of condition j being present in the school (denoted by $Q_{ij}=1$) rises along a logistic curve that bounds the probability between 0 and 1. For each test item j, the above function describes a distinct logistic Item Characteristic Curve (ICC), which graphs the probability of the condition being present as a function of θ_i.

The "difficulty" parameter for question j, b_j, is the value of θ_i at which the probability of condition j being present is 50 percent. As an item's value of b_j rises, its ICC shifts to the right, indicating greater difficulty (because higher levels of θ are required to achieve any given probability of the condition being present).

The discrimination parameter for item j is a_j. The greater is a_j, the steeper is the item's ICC for values of θ_i near the threshold b_j, and the more sensitive this item is to changes in the latent variable near the difficulty level. Higher values of a_j (while holding the difficulty parameter constant) imply that an item is more informative for distinguishing among schools with different levels of management quality near the item's difficulty level.

For a set of candidate test questions to do a good job of estimating the underlying values of the latent trait over most of the latent trait's range, the test items should have difficulty parameters spread between -2.5 and 2.5. The difficulty parameters are measured in the same "units" as the underlying latent variable, so that having difficulty parameters spanning the range [-2.5, 2.5] implies that the test has some assessment items that even schools at the lowest levels of school management quality have some probability of passing, while also containing questions that are challenging even for schools at the highest levels of school management quality. Ideal items also have discrimination parameter values that are "high enough," with different psychometricians offering different rule-of-thumb minimum thresholds, usually in the range of 0.65 to 1.0.

The model of equation (1) extends this framework to incorporate polychotomous responses in a natural way. Called the "graded response model," this model treats a categorical variable with K+1 possible values (0, 1, …, K) as if it were revealing the presence or absence of K conditions or "tests" of increasing difficulty. All threshold tests for the same question share the same discrimination parameter a_j, but differ in difficulty parameter b_{jk}. This is an attractive way to incorporate data on categorical questions regarding the frequency of school management practices and on Likkert-scale opinion questions.

Implicit in this model are two inter-related assumptions, which we will seek to evaluate empirically. First, we assume that the notion of school management quality relevant to our study is uni-dimensional. That is, we assume that a school's performance on all assessment items is tied to the school's value of a single quality trait. We articulate the notion of school management quality that we have in mind in the next section, and the assessment items we will use for revealing the level of that trait in the subsequent section. Second, we make the "local independence assumption," that a school's responses to different assessment items are independent conditional on θ_i. In our case this means that, for any school with management quality level θ_i, the responses to any two questions employed in the analysis are not correlated. This means that nothing should link responses to two questions for a given school other than the school's level of management quality. One practical implication of this assumption is that, even though we obtained data on some management practices from both head teachers and teachers, we use only the median teacher response as a candidate test item; two measures of the same practice would likely be correlated, violating this assumption. More subtly, the local independence assumption could be violated if school management practice choices are driven by more than

one latent trait (violating the uni-dimensionality assumption as well). We will further examine the local independence assumption and uni-dimensionality assumptions below.[10]

The "school management quality" construct

In the broadest sense, "school management" includes activities aimed at maximizing a school's achievement of its educational objectives, subject to the opportunities and constraints that it faces. These activities may include: 1. Gathering information to assess school needs, challenges and opportunities (including opportunities for government-provided teacher trainings); 2. Researching and brainstorming possible innovations and solutions; 3. Coordinating activities and encouraging cooperation among the schools' teachers and stakeholders; 4. Encouraging, mentoring, monitoring and rewarding individual teachers; and, more generally, 5. Providing leadership for daily activities and for school improvement efforts. Holding constant the school's resources, challenges and external opportunities (for, say, assistance from the central government), higher quality management allows the school to achieve more of its educational objectives.

While, in general, better or more intense school management activities might be directed toward some combination of increasing enrollment, increasing outreach among disadvantaged groups, and improving learning outcomes, for the purposes of the present study, we focus more narrowly on the subset of school management practices that may be directed toward improving student learning by improving teaching practices. We are, therefore, especially interested in managerial efforts related to encouraging, equipping, coordinating, monitoring, mentoring, and rewarding or punishing teachers in ways that foster greater teacher effort, care and innovation.

In principle, diverse actors might participate in these school management activities – including head teachers, School Management Committee members, Parent Teacher Association committee members, and other teachers, parents or community members – and the exact mix of actors who step forward to undertake these activities may differ from school to school. In practice, however, baseline data and earlier qualitative research suggest that head teachers are the actors most likely to undertake many of the day-to-day management activities most closely associated with teachers' pedagogical practices, such as observing teachers' classroom practices, providing feedback to teachers, and coordinating teachers' daily activities. School Management Committees, by contrast, tend to concern themselves primarily with matters of fundraising and maintaining/improving physical infrastructure, and visit classrooms only on rare occasions. Parents, too, are unlikely to play strong roles in teacher management, because of their low levels of education and low status relative to grade 9 and 10 math and science teachers. While head teachers are the most likely to engage in teacher management practices, many have heavy teaching loads, and head teachers probably differ greatly in their capacity and motivation to do these activities. Our indicators focus primarily on the intensity and quality of head teacher practices regarding management of classroom activities.

While we assume that many potentially valuable management practices will be executed by head teachers, we remain agnostic about the driving force behind the use of these practices. The motivation behind a head teacher's managerial effort may be internal to the head teacher or externally imposed by well-motivated and able SMCs, PTAs or other community members. Such actors might help create conditions for good head teacher management practices by hiring head teachers with good qualifications and attitudes, by providing them with adequate resources and decision-making latitude, and by holding them accountable for good results.

[10] The local independence assumption would be violated by construction if we took a variable with three answer categories (e.g. "never," "sometimes," "always"), used it to create two dummies (representing, e.g., the answers "sometimes" and "always"), and included both dummies in the list of candidate test items. There is no need to do this, because polychtomous IRT models handle trichotomous response variables well.

In light of the above, for the purposes of this study, we define "school management quality" to be a uni-dimensional measure for which higher values imply the school's ability to achieve better teaching and learning outcomes, given its resources and constraints. We assume that increases in school management quality are manifest in increasing intensity of practices that encourage, monitor and coordinate teachers, and in increasing expressions among teachers that suggest they feel observed, encouraged and rewarded for working hard and innovating.

Candidate management quality assessment items

Appendix Table A1 lists the questions (asked of teachers) from which we construct our candidate management quality assessment items. The items are of two broad types. Items 1 through 7 describe the frequency of activities that could be used to monitor, mentor or coordinate teachers in their teaching efforts. While in principle one might worry that, beyond some efficient level of frequency, more frequent meetings or classroom visits could increase distraction rather than teaching quality, in practice the frequencies are sufficiently low in most schools that it seems reasonable to assume that more frequent interactions of these sorts reflect more energetic, and therefore better, management efforts. Items 8 through 18 describe the strength of teacher opinions regarding their experience of other actors' coordinating, motivating and leadership activities.

In the baseline teacher questionnaire, respondents were asked to choose among five response options for the frequency questions (not at all, 1 to 4 times during the previous academic year, once a month, once a week, and more than once a week), among four categories (fully disagree, disagree, agree, strongly agree) for Likkert-scale opinion questions, and among three categories (not at all effective, somewhat effective, very effective) for opinion questions about leadership effectiveness. Because IRT model estimation can be unstable when there are too few responses in a category, we merged extreme categories with neighboring categories when they contained fewer than 5 percent of the observations. The table reports the distribution across schools of the median responses (among a school's grade 9 and 10 math and science teachers), after rounding up to the nearest integer when the medians were half way between categories.

Preference for median teacher reports over head teacher reports

For the first seven candidate management quality assessment items introduced in Appendix Table 1, it is possible to compare the median teacher responses to head teacher responses for the same schools. The comparison is not perfect, because the response category options offered to the two types of respondents were different, with the head teachers offered more categories.[11] For comparison purposes, we collapsed the head teacher categories into five categories as similar as possible to the teacher categories. We collapsed the original head teacher categories "1-2 times during the year" and "3-6 times during the year" into a single category, for comparison to the teacher category "1-4 times during the year." For both sets of respondents, this is the only category between "not at all" and "once a month." We suspect that for true frequencies on the order of 5 to 6 times per year, the options given to teachers would lead them to round up to "once a month," while the options given to head teachers would leave them more likely to report the lower frequency. We also merged head teachers' responses of "twice a month" down to "once a month." As indicated above, we also chose, when calculating teacher median responses, to round medians up to the next integer category when the median was halfway between categories. We believe all these choices would tend to raise teacher median responses relative to head teacher responses if both sets of respondents were observing the same reality and reporting in an unbiased fashion.

[11] Teachers were offered the categories: 1= none/not at all, 2=1-4 times per year, 3=once a month, 4=once a week, and 5=more than once a week. Head teachers were offered the categories: 1=none/not at all, 2=1-2 times during the year, 3=3-6 times during the year, 4=once a month, 5=twice a month, 6=once a week, and 7=more than once a week.

Despite making these choices that would tend to raise reported median teacher frequency responses relative to head teacher responses, we in fact see head teachers reporting significantly higher frequencies for their own good management practices relative to teachers' reports of those same frequencies. For example, while 68 percent of head teachers report having meetings to discuss school activities and challenges at least once a month, only 50 percent of teacher median reports are in this frequency range. Similarly, for meetings to discuss pedagogy, while only 11 percent of head teachers report having no meetings of this sort in the last academic year and 31 percent report having such meetings at least once per month, median teacher reports suggest there were no such meetings in 32 percent of schools and such meetings took place at least once per month in only 13 percent of schools.

We interpret these differences as indicating greater probable bias in head teacher than teacher frequency reports. The frequencies we consider are frequencies of actions for which the head teachers are accountable, at least in principle. They might, therefore, feel that their responses could be used to evaluate them, and might perceive the need to make themselves look especially active by over-stating their frequencies. Teachers may possibly over-state frequencies, too, if they wish to make their head teachers look good, or if they think that being visited and monitored more frequently would improve the way they appear to interviewers, but this possible motivation for biased reporting seems significantly weaker for teachers than for head teachers. Perhaps more important, we cannot think of any reason why the teachers would bias their frequency reports downward. Thus we are inclined to think that the teachers' reports are more accurate than the head teachers'. Spearman rank correlations and simple regressions show that the head teacher and median teacher reports are statistically significantly correlated, but not tightly correlated. This suggests that if the teacher median scores are more accurate, using the head teacher scores instead would tend to produce misleading rankings. We find here a strong reason to prefer the use of teacher rather than head teacher reports for measuring school management quality as defined in this study.

Evaluation and selection of final management quality assessment items

Appendix Table A2 presents descriptive statistics that shed light on the difficulty and discriminatory powers of the candidate assessment items (without the restrictions of IRT model estimation). Column A offers a crude description of the items' relative "difficulties": the mean values of the responses across all schools in the sample after transforming all variables so that their values range from 0 to 100 points. The items are ordered in increasing difficulty, as indicated by decreasing mean scores across all schools. At least by this crude measure, the items appear to span a wide range of difficulties, as is desirable.

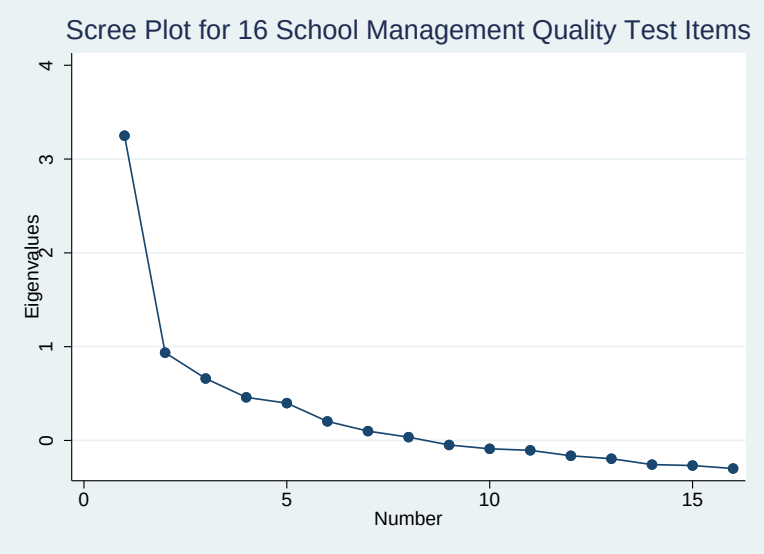
Column B presents "item-rest correlations," which are correlations between the individual indicators and "rest scores" created by summing the other 17 indicators. Assuming that the latent variable of interest is correlated with the "rest score," lower values for the "item-rest correlation" suggest items that have less discriminatory power for estimating the value of the latent variable. In most cases, these correlations are positive and moderately high, suggesting that this sets of items is coherent and rendering reasonable the assumption that they all relate to a common latent trait. The two possible exceptions are items 6 (frequency of formal, written feedback from head teacher to teachers) and 7 (frequency of visits of SMC members to classrooms), both of which are quite rare activities in sample schools. In what follows we will remove these items from the list of candidate assessment items.

Psychometricians report the Chronbach's alpha measure as a measure of the reliability of the total score across a set of assessment items as a measure of an underlying latent trait. Under the assumptions of Classical Test Theory (see Wu, et al., 2017), it may be interpreted as an estimate of the correlation between the total score on the given set of assessment items and a test composed of the same number of items randomly selected from the population of items that might be used to test competency in the same domain. It is also an estimate

of the mean correlation coefficient one would find if one split the items in the assessment into halves in all possible ways, and for each split calculated the correlation between scores on the two halves. The value of Chronbach's alpha for the 18 items included in Table 2 is 0.77. After removing items 6 and 7, the value is 0.78. While the value here is lower than the 0.8 or 0.9 that psychometricians typically hope for when developing academic achievement tests, it is nonetheless high enough to suggest a reasonable degree of coherence and reliability.

To examine the reasonableness of the uni-dimensionality assumption, psychometricians typically use Principal Components Analysis (despite this being more appropriate for continuous measures). Figure 1 shows the resulting scree plot. The variance of the first principal component, which places positive loadings on all 16 indicators, explains 70 percent of the variation in the data and its eigenvalue is much larger than for the other components. We interpret this as providing reasonably strong support for proceeding with the IRT analysis under the uni-dimensionality assumption.

Appendix Figure A1



IRT model estimates and properties

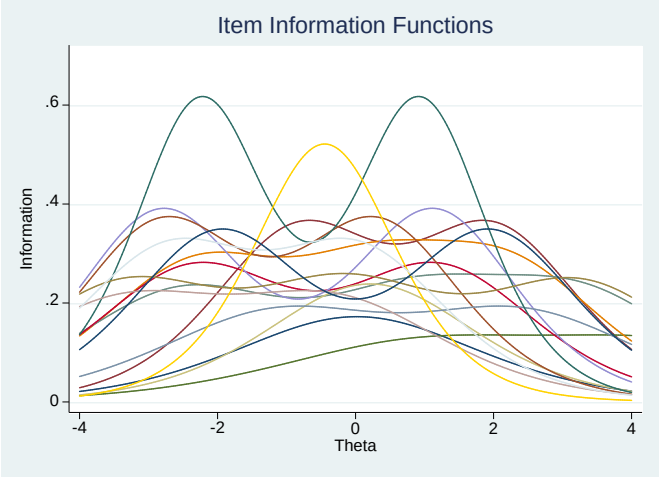
Appendix Table A3 presents estimates of the graded response model employing the 16 school management quality test items selected from the above analysis. Column A presents the estimates of the discrimination parameters (which are shared for all of the item's difficulty parameters), while columns B through D present the difficulty parameters associated with each value above the first for the polychotomous items. The items are listed in increasing order of the discrimination parameter estimate, where higher values indicate items that are more informative regarding the value of the latent variable in the ranges of the latent variable near the item's difficulty parameter value.

Rules of thumb employed in psychometrics suggest including in a final measure only items with discrimination parameters greater than 0.65, 0.8 or 1.0 (depending on the expert). While all items pass the most liberal of these tests, the discrimination parameter for item 3 (the frequency with which the head teacher observes a teacher's full class session) is borderline.

Because discrimination parameters describe an item's discrimination power only near its difficulty parameter values, it is useful to graph out more completely each item's Item Information Function (IIF), which graphs the Fisher Information measure for estimates of the latent variable across its range.

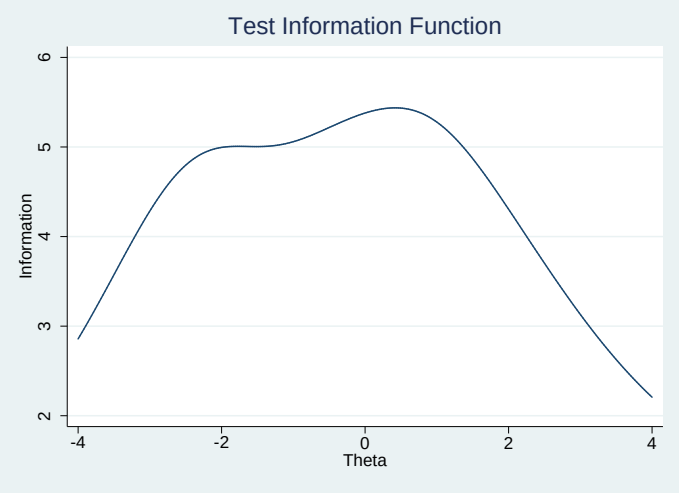
Appendix Figure A2 presents the IIFs for the 16 items included in estimation. The item with the tallest peaks to its IIF is Item 13, which indicates the strength of agreement with the statement: "When teachers experiment with using new teaching methods, the head teacher notices and encourages it."

Appendix Figure A2



Appendix Figure A3 plots the Test Information Function, which is the sum of the IIFs. It suggests that the set of items included in the estimation are informative over a broad range of latent management qualities, though somewhat less so at the upper end of the relevant range (near 2) than at the lower end.

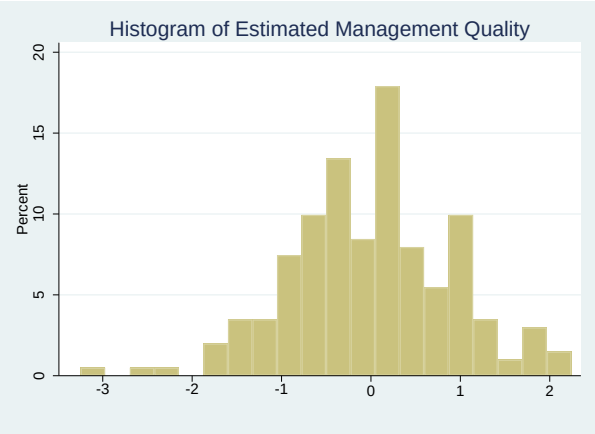
Appendix Figure A3



Management quality index values

Appendix Figure A4 presents a histogram of the values of the index (θ) implied by the estimates of Appendix Table A3 for each of the 201 schools in our sample for which we had teacher questionnaire data. We will henceforth call this estimated index "Theta1." The correlation Theta1 and a simple total of the 16 indicators included in the estimation is 0.991.

Appendix Figure A4



Correlations of management quality index with baseline determinants and outcomes

One way to assess the validity of Theta1 is to study its correlations with other variables (from baseline) for which we would expect correlation with the quality of a school's management of teachers and classroom practices. These potential correlates might include potential determinants of school management quality, indicators of the quality of school management along dimensions other than the management of teachers and classroom practices, and teaching and learning outcomes that might be improved by higher school management quality.

Appendix Table A4 present statistics describing a variety of such correlations. None of these should be interpreted as an estimate of a causal effect. For simplicity in interpreting the sizes of the associations, the correlation between each variable and Theta1 is described by running a Weighted Least Squares regression of the variable on Theta1, where the weights are population weights (weight1 as described in the main text), and with the regressions also including district and priority/non-priority stratum fixed effects. For teacher- and student-level regressions, the standard errors are also clustered at the school level. The units of Theta1 correspond roughly to standard deviations of the school management quality distribution, so the coefficients may be interpreted as indicating the size of the average increase in the potential correlate associated with a one standard deviation increase in the school management quality index.

The results in Appendix Table A4 provide some support for the validity of Theta1 as a measure of schools' quality of management for teachers and teaching practices. This measure of school management quality is statistically significantly lower in in more remote schools (i.e. schools that are farther from the nearest all-weather motorable road), as we might expect, but it is not statistically significantly correlated with the head teacher having a management degree or with how much time the head teacher must spend teaching. It is statistically significantly correlated (at least marginally) in the expected directions with some school-level outcomes that might result from good school management, namely whether many parents participated in a government mandated social audit for the school, whether the school provides free coaching for low-performing students, and (negatively) with grade 9 and 10 student reports of their teachers being frequently absent. With one exception, however, it is not statistically significantly correlated with student test scores. While these weak correlations with test scores fail to provide strong support for the validity of the management quality index, we do not believe that they refute the validity, because there are at least two other reasons why the correlations might be low (even if the index is a valid measure of school management quality): barriers to learning other than poorly motivated teachers may be so great for most students that better school management quality alone has little impact on their learning, and the baseline student assessment scores may

contain much more noise than signal (for reasons related to the quality of assessments and poor test-taking discipline). Improvements in the quality of assessments and assessment administration at endline may yet reveal greater correlation.

Table 1
Descriptive Statistics (Using Population Weight) and Balance Tests for Baseline School Characteristics

Variable	Number of observations	Mean (standard error of mean)	Standard deviation	10^{th} percentile	90^{th} percentile	Mean for TT Sample	Mean for TTVA Sample	Mean for Control Sample	p-value for test of $\beta_{TT} = \beta_{TTVA} = 0$ (1)	p-value for test of $\beta_{TT/TTVA} = 0$ (2)
School-Level Variables										
Log of total no. of students in school	203	5.91 (0.04)	0.546	5.19	6.57	5.85 (0.08)	5.86 (0.09)	5.96 (0.06)	0.318	0.130
Student/teacher ratio in grades 9 and 10	203	14.6 (0.74)	10.2	6.0	26.6	13.5 (1.2)	14.9 (1.9)	15.0 (1.0)	0.388	0.410
Share of teachers with permanent contracts	203	0.365 (0.02)	0.181	0.118	0.600	0.378 (0.025)	0.362 (0.040)	0.360 (0.020)	0.666	0.730
Share of teachers who are female	203	0.295 (0.010)	0.128	0.143	0.500	0.287 (0.021)	0.306 (0.019)	0.294 (0.013)	0.821	0.924
Days school was open last year (grade 9)	203	195.6 (1.29)	16.2	180	218	198.8 (2.7)	191.1 (2.8)	196.3 (1.6)	0.059*	0.574
Hours to nearest all-weather road	203	3.16 (0.36)	4.54	0.05	8.00	3.27 (0.74)	2.45 (0.41)	3.47 (0.58)	0.468	0.427
Head teacher has a manage-ment degree	203	0.585 (0.039)	0.494	0	1	0.576 (0.078)	0.629 (0.076)	0.567 (0.055)	0.813	0.678
Hours per week that head teacher teaches	203	16.38 (0.51)	6.82	8.25	25.50	17.27 (0.88)	16.65 (1.05)	15.81 (0.75)	0.455	0.238

School has electricity (most days)	203	0.771 (0.032)	0.421	0	1	0.808 (0.063)	0.810 (0.058)	0.732 (0.048)	0.351	0.149
Students take assessments more frequently than trimestral	203	0.850 (0.028)	0.358	0	1	0.872 (0.050)	0.899 (0.044)	0.815 (0.045)	0.431	0.237
Share of students reporting math teacher absent frequently or very frequently	203	0.063 (0.004)	0.052	0	0.143	0.065 (0.008)	0.072 (0.009)	0.057 (0.005)	0.244	0.126
Share of students reporting science teacher absent frequently or very frequently	203	0.070 (0.004)	0.054	0	0.151	0.072 (0.008)	0.078 (0.009)	0.065 (0.006)	0.504	0.277
Estimated management quality index	201	-0.015 (0.072)	0.898	-1.116	1.092	-0.085 (0.168)	0.167 (0.111)	-0.075 (101)	0.246	0.483
Teacher-level variables										
Teacher has at least bachelor degree in math or science	393	0.740 (0.025)	0.439	0	1	0.667 (0.055)	0.741 (0.046)	0.774 (0.034)	0.183	0.139
Hours per day that teacher preps for class	393	0.807 (0.062)	0.964	0.083	2.000	0.676 (0.097)	0.892 (0.136)	0.823 (0.086)	0.284	0.551
Teacher had SSRP training	393	0.333 (0.033)	0.517	0	1	0.254 (0.060)	0.367 (0.062)	0.351 (0.049)	0.347	0.445

Student-level variables										
Father has secondary or higher education	16,435	0.264 (0.01)	0.441	0	1	0.272 (0.031)	0.254 (0.014)	0.266 (0.013)	0.916	0.979
Mother has secondary or higher education	16,435	0.110 (0.01)	0.313	0	1	0.093 (0.017)	0.115 (0.012)	0.115 (0.011)	0.279	0.352
Number of assets owned by family (scooter, bike, computer, TV, refrig.)	16,431	1.26 (0.08)	1.23	0	3	1.16 (0.11)	1.37 (0.19)	1.25 (0.11)	0.559	0.855
Math score (IRT), grade 8	7,651	0.011 (0.045)	0.864	-1.033	1.150	-0.002 (0.096)	0.017 (0.103)	0.014 (0.055)	0.685	0.419
Science score (IRT), grade 8	7,651	0.054 (0.059)	0.907	-1.113	1.203	0.042 (0.146)	0.110 (0.141)	0.030 (0.058)	0.709	0.542
Math score (IRT), grade 9	8,784	0.016 (0.053)	0.918	-1.115	1.259	-0.019 (0.090)	-0.008 (0.136)	0.043 (0.068)	0.476	0.224
Science score (IRT), grade 9	8,784	0.016 (0.046)	0.844	-1.069	1.109	-0.040 (0.072)	0.011 (0.101)	0.043 (0.066)	0.479	0.400

(1) p-value from test of hypothesis that coefficients on TT and TTVA are zero in WLS regression of variable on TT TTVA and district and priority stratum fixed effects, with weight equal to weight1.

(2) p-value from test of hypothesis that coefficient on Treat (=TT+TTVA) is zero in WLS regression of variable on Treat and district and priority stratum fixed effects, with weight equal to weight1.

Appendix Table A1
Candidate Management Quality Index Items

Item No.	Item Name	Question	Value=Definition	Percentage Frequencies
Activity frequencies				
1	General teacher meetings	Please indicate the frequency during the last academic year of teachers' gathering (with or without head teacher) to discuss school activities or challenges	0= 0-4 times per year 1= At least once a month	50.3 49.7
2	Pedagogy meetings	Please indicate the frequency during the last academic year of teachers' gathering (with or without head teacher) to learn about a new teaching method or new academic materials	0= Not at all 1= 1-4 times per year 2= At least once a month	31.8 55.2 12.9
3	HT observes full class	Please indicate the frequency during the last academic year of observation by the head teacher of a full class period	0= Not at all 1= 1-4 times per year 2=At least once a month	65.2 29.4 5.5
4	HT observes part of class	Please indicate the frequency during the last academic year of observation by the head teacher for a short duration (less than a full class period)	0=Not at all 1= 1-4 times per year 2= Once a month 3= At least once a week	11.4 44.8 31.3 12.4
5	HT gives informal feedback	Please indicate the frequency during the last academic year of informal, verbal feedback from your head teacher on your teaching	0= Not at all 1= 1-4 times per year 2= Once a month 3=At least once a week	11.4 55.7 26.4 6.5
6	HT gives formal feedback	Please indicate the frequency during the last academic year of formal, written feedback from your head teacher on your teaching	0= Not at all 1= At least once per year	94.0 6.0
7	SMC visits to classrooms	Please indicate the frequency during the last academic year of observation of	0=not at all 1=at least once per year	61.7 38.3

		your class by a member of the School Management Committee who is not also a teacher or head teacher in the school		
Teacher opinions [Please indicate the extent to which you agree or disagree with the following statement:]				
8	Growth opportunities	Please indicate the extent to which you agree or disagree: "The school has provided ample opportunities for me to learn and become a better teacher."	0=Strongly disagree/disagree 1=Agree 2=Strongly agree	11.4 62.2 26.4
9	Team work encouraged	Please indicate the extent to which you agree or disagree: "The school encourages teachers to work as a team."	0=Strongly disagree/disagree 1=Agree 2=Strongly agree	5.5 69.7 24.9
10	Cooperation helps me	Please indicate the extent to which you agree or disagree: "Cooperation and knowledge exchange with other teachers helps me become a better teacher."	0=Less than strongly agree 2=Strongly agree	54.2 45.8
11	HT encourages hard work	Please indicate the extent to which you agree or disagree: "When teachers work hard, the head teacher notices and encourages it."	0=Strongly disagree/disagree 1=Agree 2=Strongly agree	6.0 51.2 42.8
12	Parents encourage hard work	Please indicate the extent to which you agree or disagree: "When teachers work hard, parents notice and encourage it."	0=Strongly disagree/disagree 1=Agree 2=Strongly agree	30.4 55.2 14.4
13	HT encourages innovation	Please indicate the extent to which you agree or disagree: "When teachers experiment with using new teaching methods, the head teacher notices and encourages it."	0=Strongly disagree/disagree 1=Agree 2=Strongly agree	7.0 66.7 26.4
14	Parents encourage innovation	Please indicate the extent to which you agree or disagree: "When teachers experiment with using new teaching methods, parents notice and encourage it."	0= Strongly disagree `1= Disagree 1=Agree 2=Strongly agree	5.5 40.8 47.8 6.0

15	Enjoy teaching here	Please indicate the extent to which you agree or disagree: “I enjoy teaching in this school.”	0=Strongly disagree/ disagree 1=Agree 2=Strongly agree	6.5 36.3 57.2
16	Head teacher leadership	In your opinion, how effective is the head teacher in providing leadership for the school?	0=Not or somewhat effective 1=Very effective	38.3 61.7
17	SMC head leadership	In your opinion, how effective is the School Management Committee head in providing leadership for the school?	0=Not effective 1= Somewhat effective 2= Very effective	7.5 42.3 50.3
18	SMC member leadership	In your opinion, how effective are other School Management Committee members in providing leadership for the school?	0=Not effective 1= Somewhat effective 2= Very effective	13.4 73.1 13.4

Appendix Table A2
Descriptive Statistics for Management Quality Assessments Items, In Increasing Order of Difficulty

Indicator Number	Indicator Label	A. Mean Value on 0 to 100 Scale	B. Item-Rest Correlation
ind15	I enjoy teaching here	75.4	0.34
ind17	SMC head leadership	71.4	0.39
ind11	HT encourages hard work	68.4	0.37
ind16	Head teacher leadership	61.7	0.43
ind09	Team work encouraged	59.7	0.41
ind13	HT encourages innovation	59.7	0.47
ind08	Growth opportunities	57.5	0.37
ind14	Parents encourage innovation	51.4	0.35
ind18	SMC member leadership	50.0	0.41
ind01	General teacher meetings	49.8	0.32
ind04	HT observes part class	48.3	0.44
ind10	Cooperation helps me	45.8	0.32
ind05	HT gives informal feedback	42.6	0.39
ind12	Parents encourage hard work	42.0	0.33
ind02	Pedagogy meetings	40.5	0.47
ind07	SMC visits classrooms	38.3	0.18
ind03	HT observes full class	20.1	0.30
ind06	HT gives formal feedback	6.0	0.17

Appendix Table A3
Estimates of Graded Response Model (Standard errors in parentheses)

ind	label	A. Discrimination parameter	B. Difficulty 1	C. Difficulty 2	D. Difficulty 3
ind03	HT observes full class	0.704 (0.201)	0.976 (0.325)	4.332 (1.190)	
ind01	General teacher meetings	0.834 (0.211)	0.018 (0.195)		
ind12	Parents encourage hard work	0.859 (0.191)	-1.100 (0.286)	2.361 (0.497)	
ind15	I enjoy teaching here	0.917 (0.205)	-3.301 (0.684)	-0.355 (0.191)	
ind05	HT gives informal feedback	0.957 (0.195)	-2.471 (0.473)	0.898 (0.229)	3.167 (0.618)
ind10	Cooperation helps me	0.978 (0.229)	0.212 (0.177)		
ind14	Parents encourage innovation	0.985 (0.197)	-3.303 (0.627)	-0.143 (0.174)	3.227 (0.607)
ind08	Growth opportunities	1.051 (0.208)	-2.312 (0.420)	1.184 (0.251)	
ind04	HT observes part class	1.066 (0.197)	-2.275 (0.399)	0.284 (0.166)	2.171 (0.381)
ind17	SMC head leadership	1.121 (0.223)	-2.676 (0.481)	-0.010 (0.157)	
ind18	SMC member leadership	1.177 (0.239)	-1.962 (0.348)	1.950 (0.344)	
ind11	HT encourages hard work	1.211 (0.229)	-2.783 (0.466)	0.304 (0.156)	
ind09	Team work encouraged	1.248 (0.241)	-2.800 (0.470)	1.135 (0.219)	
ind16	Head teacher leadership	1.444 (0.308)	-0.447 (0.149)		
ind13	HT encourages innovation	1.567 (0.282)	-2.235 (0.326)	0.929 (0.171)	
ind02	Pedagogy meetings	1.912 (0.230)	-0.805 (0.194)	1.980 (0.340)	

Appendix Table A4
Results of Weighted Least Squares Regressions of Possible Correlates on Theta1*

Potential Correlate	Number of Observations	Estimated Slope on Theta1 (Std. Error)	p-value for two-tailed test of β=0
School-level variables			
Inverse Hyperbolic Sine Transformation of Hours to Walk to the Nearest All-Weather Road	201	-0.162 (.082)	0.049
Whether Head Teacher has a management degree	201	-0.031 (.049)	0.535
Hours per week that the HT teaches	201	-0.160 (.561)	0.775
Whether HT reports constant pressure from most parents to achieve high academic standards	201	0.047 (.030)	0.117
Whether the school performed a social audit in which at least half the parents participated	201	0.098 (.040)	0.015
Whether school gives prizes for high-performing students	201	0.053 (.042)	0.214
Whether school provides free coaching for low-performing students	201	0.060 (.035)	0.085
Fraction of students reporting math teacher absent frequently or very frequently	201	-0.014 (.004)	0.001
Fraction of students reporting science teacher absent frequently or very frequently	201	-0.009 (.005)	0.063
Teacher-level variables**			
Hours teacher spends preparing per class	393	-.082 (.074)	0.269
Student-level variables**			
Math score (IRT), grade 8	7596	0.052 (.042)	0.214
Science score (IRT), grade 8	7596	0.022 (.048)	0.648
Math score (IRT), grade 9	8727	0.050 (0.43)	0.247
Science score (IRT), grade 9	8727	0.051 (0.040)	0.205

* Regressions include district and priority/non-priority stratum fixed effects, and employ weight1 as described in text.

** For regressions using teacher- or student-level data, the standard errors are clustered at the school level.