Enhancing Teacher Productivity with Information on Students and Skills: Experimental Evidence from Bangladesh*

Sharnic Djaker† University of Oxford
Alejandro J. Ganimian‡ New York University
Noam Angrist§ University of Oxford
Shwetlena Sabarwal¶ World Bank

September 16, 2026

Abstract

Teachers may fail to address learning gaps because they do not know which students and skills need support. While exams are prevalent in every education system, these tests often go underutilized. We evaluate an intervention that provided grade-6 math teachers in 312 schools in Bangladesh with results from two brief tests, reports identifying topics and students in need of support, and an annotated answer key. At baseline, teachers could not estimate their students' test scores, class rank, or spread, and they were twice as likely to interact with high- than with low-achieving students. The reports redirected instruction toward the needs they identified: flagged topics and students improved more on the intervention assessments. On independent endline tests, the program raised math achievement by 0.23 standard deviations (SDs) and reading by 0.17 SDs, with similar gains across the baseline distribution. It also improved three social-emotional skills by 0.10-0.14 SDs. Teachers lowered their estimates of students' performance, distributed interactions more equitably, and emphasized more foundational content. At $5.06 per student, these findings show that actionable assessment results utilizing exam data in education systems can help teachers become substantially more productive at low cost.

*We thank the Foreign, Commonwealth & Development Office, the World Bank, the National Academy for Education, the Spencer Foundation, and the Jacobs Foundation for the funding for this study. We also thank Hunter Gelbach for his contributions at various stages of this project. Seminar participants at Chicago, Harvard, NYU, and ProLeer provided comments on earlier drafts.

†Senior Research Scientist, What Works Hub for Global Education, Blavatnik School of Government, University of Oxford. sharnic.djaker@bsg.ox.ac.uk.

‡Associate Professor of Applied Psychology and Economics, Steinhardt School of Culture, Education, and Human Development, New York University. alejandro.ganimian@nyu.edu.

§Academic Director, What Works Hub for Global Education, Blavatnik School of Government, University of Oxford. noam.angrist@economics.ox.ac.uk.

¶Lead Economist, Education, World Bank. E-mail: ssabarwal@worldbank.org.

1 Introduction

A primary vehicle governments rely on to foster human capital is student-teacher interaction in school. Yet, several information frictions may constrain the productivity of such interactions. First, teachers may fail to know which students need the most support; second, teachers may also not know which skills require attention. While substantial information often exists on students and skills through regular exams in nearly every education system worldwide, these tests often go underutilized. This gap is consequential, introducing large information frictions which blunt the productivity of teacher-student interactions, but could be addressed at low cost and at scale by better leveraging existing exams in the system. Closing these types of information frictions has proved important in the context of parent-child interactions (Bergman, 2020; Dizon-Ross, 2019) and could be even more essential in the context of teacher-student interactions, helping explain why despite historic expansions in schooling worldwide, student learning levels remain low and stagnant (Angrist et al., 2021; World Bank, 2018a).

In this paper, we experimentally evaluate an intervention designed to close information frictions along two margins, identifying *students* and *skills* in need of support, by providing grade-6 math teachers in Bangladesh across 312 schools with actionable results from tests. We drew a representative sample of 312 public-private partnership (PPP) secondary schools—which enroll 8 in 10 students—and randomly assigned them to a treatment or a control group. In the former, we assessed grade-6 students at the start and middle of the year and, shortly thereafter, offered teachers reports identifying the topics and students in need of support. Teachers also received the test and an explanation of why each answer was correct or incorrect to reinforce key concepts and identify the misconceptions underlying frequent student errors. Control schools did not participate in the assessments or receive the reports and error guides, so that we could evaluate their impact relative to business-as-usual instruction and learning.

We report results in three main segments. First, we quantify information frictions among teachers and their consequences for instruction. Teachers could not estimate their students' performance on tests. We tested students in math and asked teachers to estimate their scores. The correlation between actual and estimated scores was 0.23, so estimates explained only 5% of test-score variation. Teachers did not seem to know which students needed more support: the correlation between students' actual within-class ranks and those implied by estimates was 0.16, which means teachers' rankings predicted less than 3% of the variation in actual rankings. The magnitudes of their errors were large relative to the class distribution: they overestimated the scores of their lowest-achieving students—who most need their support—by 6.9 percentage points (pp.), which is equivalent to 50% of the within-class standard deviation (SD) in scores. Teachers did not even know how much students in their class varied in their achievement: the correlation between the within-class SDs of actual scores and teachers' estimates was -0.01.

These results suggest that teachers may be insufficiently catering to the needs of their students partly because they are unaware of how many are struggling or the extent to which they do.

Teachers had few opportunities to update their incorrect priors about their students' skills. We observed lessons and collected detailed data on every verbal exchange a teacher had with an individual student or small group. In the typical 40-minute math lesson, there were only 8.5 such interactions; most of the lesson was devoted to whole-group instruction. The average student interacted just 0.24 times with their teacher, but even those scarce exchanges were unequally distributed: the lowest-achieving students interacted half as often as their highest-achieving peers; in fact, 9 in 10 low achievers did not interact with their teachers at all. These differences were driven by teachers (rather than students), who initiated 7 in 10 exchanges.

Second, we report the effects of the intervention, which was designed to close these gaps through information on both students and skills. Teachers seemed to respond to the reports. In treatment schools, students improved 14.5 pp. more on topics (e.g., measurement and geometry) flagged for low performance than on non-flagged topics. They also improved more on flagged sub-topics (e.g., applying geometric properties to find a missing angle) by 19.4 pp. than on non-flagged sub-topics. Even the color with which we displayed item-level performance appeared to make a difference: students improved 34.7 pp. more on items shown in red (low performance) or yellow (medium) than on items shown in green (high). Students mentioned by name in the reports also improved 9.9 pp. more, and the magnitude of their gains increased with the number of topics in which they were flagged for low performance. These students even improved 32.9 pp. more in the topics for which they were flagged. These patterns imply that teachers shifted instruction to address the specific challenges highlighted by the reports.

The intervention raised achievement beyond the tests used to produce the reports. We compared treatment and control students on independent endline tests and found that the intervention increased math achievement by 0.23 SDs, impacting all topics and skills assessed. Gains were broad-based, with impacts of similar magnitudes along the baseline distribution. Students named in the reports outperformed their control peers who would have been named, but so did non-named students by similar margins. The intervention also increased reading test scores by 0.17 SDs, even though it focused on math, and we cannot rule out that effects were equal across the achievement distribution. Gains were concentrated on vocabulary and comprehension items—two skills exercised in word problems and applicable across subjects. Taken together, these impacts show that the instructional changes that teachers adopted in response to the reports benefited all students in their classrooms, not just low achievers.

The intervention also impacted some measures of students' effort and motivation. It had null effects on attendance, as measured by both school registers and the share of enrolled students present on data-collection days. Yet, the average treatment student outperformed their control counterpart on self-efficacy (0.14 SDs), growth mindset (0.11 SDs), and task-goal orientation (0.10 SDs). Low-achieving students saw statistically significant gains in register-based attendance and all three social-emotional measures. Yet, with two exceptions, we cannot rule out that effects were equal to those for their middle- and high-achieving peers. We see these effects as further evidence that the intervention produced widespread gains.

Third, in terms of mechanisms, we find larger shifts in the skills teachers teach (to the whole class) rather than targeting of specific students. Teachers adjusted their estimates of the average student's test score downward by 0.47 SDs, and we cannot reject that they adjusted by a similar magnitude for students along the achievement distribution.[1] Consistent with this uniform shift in their estimations, teachers began to distribute interactions more equitably: low achievers' share rose by 5 pp. (from 22% to 27%), mirroring the 5 pp. decline among high achievers (from 45% to 40%). Accordingly, they also shifted instruction from more advanced to more foundational topics: during observations, treatment teachers were 15.3 pp. less likely than control teachers to cover data and 13.6 pp. more likely to cover number.[2] Together, these changes help explain the generalized impacts on academic and social-emotional skills.

Our first contribution is to the literature on information. Prior work shows that parents often have inaccurate beliefs about their children's skills and that correcting those beliefs can change their investments and improve learning outcomes (Neilson et al., 2015; Bergman, 2020; Dizon-Ross, 2019; Evans and Acosta, 2026). We find that teachers also hold inaccurate beliefs about their students' achievement and, like parents, under-utilize information already available. Previous studies have also cautioned that information alone may be insufficient to produce lasting changes in teachers' behavior (Muralidharan and Sundararaman, 2010). Our study—when read alongside others (Andrabi et al., 2017; Camargo et al., 2018; de Hoyos et al., 2021; de Hoyos et al., 2023; Cilliers and Habyarimana, 2026)—suggests that when information is used to shift what the school system expects of teachers, it is more likely to raise their effort. We also document *how* this response occurs in the classroom: teachers shifted instruction towards the topics identified as needing support, improving student achievement at low cost. The effects we find from sharing assessment results without prescribing a pedagogical approach suggest that information can make the autonomy teachers already have more productive (Aghion and Tirole, 1997), which is consistent with mounting evidence from education in LMICs (Beg et al., 2022, 2024; Piza et al., 2024; Andrabi et al., 2025; Nourani et al., 2025).

Our second contribution is to a key question at the intersection of development and education of whether teachers in LMICs "teach to the top" of the achievement distribution. The concern that teachers may cater to their best students at the expense of their peers (Banerjee and Duflo, 2011) motivated some of the most influential evaluations in education, including remedial instruction (Banerjee et al., 2007), ability tracking (Duflo et al., 2011),

[1] The reports did not make teachers more accurate about the scores of individual students, which makes sense because they identified students and topics needing support rather than providing scores.

[2] Estimates on these measures are less precise because they are based on a subsample of classrooms.

learning camps (Banerjee et al., 2010), and differentiated instruction (Banerjee et al., 2017). Yet, it has been largely inferred from system incentives—such as overly ambitious curricula (Pritchett and Beatty, 2015) and high-stakes exams (Glewwe et al., 2009)—and the wide gap between curricular standards and students' preparation for school (Muralidharan et al., 2019). We are some of the first to provide *direct* evidence that business-as-usual instruction indeed focuses on advanced topics and high-achieving students.[3] Our results suggest that instruction may be mismatched to students' needs also because teachers are unaware of their achievement. Low-cost information based on student assessments can shift instruction and improve teacher productivity without broader changes to education-system structures or incentives.[4]

Relatedly, our third contribution is to research on differentiated instruction at school—one of several interventions under the umbrella term "Teaching at the Right Level" (TaRL). Randomized evaluations in India, Kenya, Ghana, Côte d'Ivoire, and Zambia have found that, when it is implemented with fidelity, this approach—which combines assessing students, grouping them by performance, and assigning different activities to each group—improves foundational skills (Duflo et al., 2011; Banerjee et al., 2017; Duflo et al., 2024; de Barros and Lubozha, 2026; Wolf et al., 2026). Our intervention sheds light on one of its mechanisms. Providing teachers with assessment results—without grouping students or assigning separate activities—led them to adjust their beliefs about students' achievement downward, distribute interactions more equitably, and teach more foundational topics, while raising learning across the baseline distribution. These findings suggest that updating teachers' beliefs may account for a non-trivial share of the effects of differentiated instruction by changing how teachers allocate attention and pitch classroom instruction for academically diverse groups of students.

Finally, we contribute to research on how to improve human capital cost-effectively in LMICs—a top priority given persistently low learning levels worldwide (Kremer et al., 2013; World Bank, 2018b; Angrist et al., 2021). Although school exams are prevalent in nearly every education system, their results often go underutilized. Our findings suggest that converting assessment results into actionable information for teachers could improve learning more cost-effectively than leading alternatives. Our intervention generated 5.68 Learning-Adjusted Years of Schooling (LAYS) in math and 4.20 LAYS in reading per $100 spent, exceeding estimates for two widely implemented approaches in global education: structured pedagogy and TaRL (GEEAP, 2020; Angrist et al., 2025). These results suggest that education systems could make better use of existing assessments to improve learning cost-effectively at scale.

[3] Our observation-based measures build on prior instruments showing that teachers devote most of the lesson to whole-group instruction and use the same materials with all students (Abadzi, 2009; Bhattacharjea et al., 2011; Bruns and Luque, 2014; Sankar and Linden, 2014; Stallings et al., 2014; World Bank, 2017).

[4] The correlation between estimated and actual test scores is consistent with our earlier work in India and Bangladesh (Djaker et al., 2024) and with replications in Benin, Nigeria, and The Gambia (Rodriguez-Segura, 2024), and is one-eighth as large as in high-income countries (Hoge and Coladarci, 1989; Südkamp et al., 2012; Kaufmann, 2024), which suggests that this possibility warrants further consideration.

2 Experiment

2.1 Context

We conducted this study in Bangladesh, a lower-middle-income country in South Asia whose population is comparable in size to those of Brazil, Russia, and Ethiopia (World Bank, 2026). In 2021, its ministry of education requested support from the World Bank to develop a formative-assessment intervention that could be rolled out nationwide. We embedded a pilot of that intervention within a large-scale randomized evaluation to estimate its impact and identify any adjustments that may be needed. Given that Bangladesh has the third-largest secondary school system in South Asia—enrolling 15.1 million students (UIS, 2026)—we expect the lessons from our study to be helpful both for this country and its neighbors in the region.

As in many LMICs, Bangladesh has rapidly expanded access to schooling in recent decades. Primary education runs from grades 1 to 5 and is compulsory; secondary education runs from grades 6 to 12 and is not (NCTB, 2013). Net enrollment in lower secondary has more than tripled since 1990, rising to 72% (BANBEIS, 2023). Enrollments have outpaced the country's capacity to build schools and hire teachers, so the rapid pace of expansion has resulted in larger and more academically diverse classrooms. The number of secondary-school students per teacher rose from 27 in 1990 to 35 in 2018 (UIS, 2026). And 8 in 10 students in secondary school are first-generation learners—i.e., they have parents who have not completed this level (Barro and Lee, 2026), and are thus less well-equipped to help their children to navigate it. Schools are under-resourced to support these students: per-pupil secondary spending is USD 371, far below the South Asian (USD 915) and LMIC (USD 924) averages (UNESCO, 2023).

In light of this context, it is perhaps unsurprising that learning outcomes remain low and unequal. In 2014, only 40% of grade-5 students could solve two-digit subtraction problems with carrying and 31% could solve three-digit-by-one-digit division problems (BRAC, 2015). Further, just 34% of students whose mothers had no schooling earned a grade point average of 4 or above, compared to 80% of those whose mothers had 10 or more years of schooling.

Several features of the Bangladeshi system may give teachers reasons to focus on higher-achieving students. First, because secondary school is not compulsory, teachers may anticipate that lower-achieving students are more likely to drop out and under-invest in them (Banerjee and Duflo, 2011; Duflo et al., 2011). Second, until recently, the curriculum resembled that of other LMICs in its emphasis on factual recall (NCTB, 2021), potentially leaving limited time for teachers to check whether students understood the material (Glewwe et al., 2009; Banerjee and Duflo, 2011; Muralidharan and Zieleniak, 2014; Pritchett and Beatty, 2015; Muralidharan et al., 2019; Rodriguez-Segura and Mbiti, 2022). Third, grade-6 students take the National Assessment of Secondary Students (NASS)—an assessment of math, Bangla, and

English (DSHE, 2019)—which may lead their teachers to invest more effort in the students more likely to pass the exams (Glewwe et al., 2010; Duflo et al., 2011; Gilligan et al., 2022).

2.2 Sampling

We sampled 312 public-private partnership (PPP) secondary schools across three divisions of Bangladesh: Chattogram, Mymensingh, and Rajshahi.[5] We focused on PPP schools because they enroll 8 in 10 secondary-school students (DSHE, 2019). We drew our sample as follows. First, we randomly sampled three of the country's eight divisions, stratifying our selection by terciles of the divisions' average NASS scores (i.e., one division per tercile). In the next step, we randomly selected four (from an average of eight) districts per division, stratifying by the median of districts' average NASS scores (i.e., two districts above and two districts below the median). Next, we excluded non-PPP schools and Madrasas (which together account for 19% of secondary enrollment) and schools with very small (below the 5th percentile) or very large (above the 75th percentile) grade-6 classes. Lastly, among the remaining schools, we sampled 26 per district (i.e., 312 in total), and among sampled schools, we randomly selected one grade-6 section per school, and up to 20 students per section, for a total of 5,935 students. These students constitute our baseline evaluation sample, which we sought to follow across all rounds of data collection. Students in treatment schools also took the intervention assessments and could thus appear by name in the reports for teachers.

In-sample schools performed broadly similarly to out-of-sample schools on NASS math and Bangla scores, both nationally and within the sampled divisions (Table A.1 in Appendix A). This suggests that the lessons that we draw from this experiment are likely to extend not only to the approximately 3.9 million secondary-school students enrolled in the sampled divisions, but also to the approximately 9.4 million secondary-school students enrolled nationally.

2.3 Randomization

We randomly assigned the 312 schools to a formative-assessment intervention described below or a business-as-usual control group. We stratified our randomization by district.[6]

Random assignment created comparable groups of students on most characteristics. The average control student was 12 years old; 61% were female, 94% spoke Bangla at home, and their average math, reading, and fluid-intelligence scores were standardized to zero. Treatment students were slightly more likely to be female, less likely to have a mother who completed high school, had higher household assets, and lower math achievement (Table A.2). We suspect

[5] Appendix Figure A.1 shows the geographic distribution of sampled schools.

[6] This study was conducted as part of an institutional partnership between the World Bank and the Government of Bangladesh, under which we also studied an intervention in which teachers were also asked to assess students. The result of this intervention is reported in a companion paper (Djaker, 2023).

that the difference in math achievement reflects the differential order in which tests were administered at baseline across groups, rather than underlying differences on unobservables.[7] Yet, to the extent that these baseline differences reflect real differences in prior achievement, they suggest that treatment students started somewhat behind control students; if anything, this would make unadjusted estimates of program impacts conservative.

Random assignment also created similar groups of teachers. The average control teacher was 42 years old and had 14 years of teaching experience; 12% were female; treatment teachers were less likely to receive pre-service training (Table A.3). Given that we conduct 14 teacher balance tests, one statistically significant difference is about what we would expect by chance. A joint test of all listed teacher covariates does not reject equality across groups.

2.4 Attrition

Most (77%) students who participated in the baseline round of data collection also participated in the endline. Attrition rates were nearly identical across experimental groups, and we find no evidence of differential attrition with respect to baseline characteristics (Table A.4).

2.5 Intervention

The intervention consisted of three components: two short (i.e., 10-item multiple-choice) math tests over eight months; two brief (i.e., five-page) reports based on those results, delivered to teachers shortly after each testing round; and a guide that identified the correct answer for each item, explained how to arrive at that answer, and described misconceptions that may explain why students chose wrong answers. Figure A.3 shows the timing of intervention activities (bottom panel).[8] Appendix B describes each component of the intervention in detail.

There are several aspects of the design of the intervention that sought to make it useful for teachers. First, we kept the tests brief so that they would be easy and low cost to administer and teachers could familiarize themselves with all items in the test and the desired process to answer them. We also expected this to make the intervention easier to scale, if successful. This short length required that we choose the topics and skills to be assessed purposefully, so we based our allocation of items on the assessment framework for an international math test for grade 4, assuming that many grade 6 students will likely be performing below grade level.[9] Second, we administered two rounds of tests, with the second round months before the end of the year to give them an incentive to improve from the first to the second round

[7]Only treatment students took the intervention tests, which preceded the impact-evaluation tests.

[8]Because of COVID-19-induced school closures, the school year started on February 22 instead of January 21, as it was otherwise intended.

[9]This assumption was based on both prior studies in LMICs (Muralidharan et al., 2019; Büchel et al., 2020) and on the results of the NASS (see section 2.1). We then checked that the topics, skills, and items assessed were aligned with the Bangladeshi curriculum and solicited feedback from Teach for Bangladesh teachers.

and to demonstrate to them that low-achieving students were indeed capable of improving.[10] Third, the reports identified the specific topics and students (by name) in need of support to focus teachers' attention on them.[11] In prior research in India and Bangladesh, we had found that teachers underestimated the share of their students that lagged curricular expectations and the extent to which they did so (Djaker et al., 2024). We wanted to understand whether making low-achieving students more visible would make teachers more willing to help them.[12] Fourth, the results were presented in narrative fashion to make them easier to understand.[13] Lastly, the guide not only identified the correct answers, but also explained how to arrive at them, and what students might be misunderstanding when they select each incorrect option (purposefully designed to be a distractor) to make the test results informative for instruction.[14]

These features map to a simple but coherent theory of change: the tests were intended to focus attention on foundational skills; the reports were meant to prompt teachers to update their priors about the topics and students in greatest need of support; the guide was supposed to translate that awareness into instructional choices; and the time between rounds gave teachers time and an incentive to improve the performance of their lowest-achieving students.

3 Data

We conducted three rounds of data collection during the 2022 school year: a pre-intervention baseline (including student and teacher surveys, student assessments, and student attendance data), a midline halfway through the year (including unannounced school visits to measure students' attendance), and an endline at the end of the year (including expanded student and teacher surveys, student assessments, classroom observations, and administrative data on students' attendance). Figure A.3 shows the timing of research activities (top panel).

3.1 Student and teacher attendance

We measure student and teacher attendance using school registers. At midline, enumerators transcribed each sampled student's and the selected math teacher's total working and present

[10] These objectives seemed important because, in a recent survey of teachers in 11 LMICs, many reported that they did not believe that they could help students if their parents were not involved in their education, and that they did not think it was their responsibility to teach lower-grade material (Sabarwal et al., 2022).

[11] We describe how topics and students were flagged in detail in section 5.2.1.

[12] This approach differs from other information interventions in education, which benchmark schools' performance against that of others in the area (Muralidharan and Sundararaman, 2010; Andrabi et al., 2017). The approach in Bangladesh was inspired by encouraging evidence from Argentina, which showed encouraging results from reports that combined such benchmarking with information on the topics in need of reinforcement (de Hoyos et al., 2021; de Hoyos et al., 2023).

[13] In a prior project, we had noticed that teachers struggled to interpret the graphs and responded more positively to reports presented in narrative form (Banerjee et al., 2018).

[14] Prior work has found that teachers in LMICs struggle with the foundational skills they are expected to teach (Tatto et al., 2012; Bruns and Luque, 2014; Bold et al., 2017; Ganimian et al., 2026).

days for January-July. At endline, they collected the same measures for August-October. We calculate attendance as the total number of days present divided by total working days across months with valid records. These data cover 5,941 students and 305 teachers at midline and 7,039 students and 311 teachers at endline.

3.2 Student characteristics and social-emotional skills

We surveyed students at baseline and endline. At baseline, we collected demographic and socioeconomic variables that we use to describe the sample, assess balance across experimental groups, and test for heterogeneous effects. At baseline and endline, we also measured students' social-emotional skills to adjust for pre-existing differences, test whether achievement impacts varied by these skills, and estimate impacts on the skills themselves.

3.3 Teacher characteristics and beliefs

We also surveyed teachers at baseline and endline. At baseline, we collected demographic, educational, and professional variables to describe the teacher sample, assess balance, and test for heterogeneous effects by teacher background. At baseline and endline, we also collected measures of teachers' beliefs and practices to adjust for pre-existing differences and test heterogeneity, and estimate impacts on these outcomes.

3.4 Student achievement

We assessed students at baseline and endline. At baseline, we measured math and reading achievement and fluid intelligence to describe the sample, assess balance, adjust for pre-existing differences, and test for heterogeneous effects by prior skills. At endline, we assessed math to estimate impacts and reading to test for spillovers. The math tests allocated items across topics (number, measurement and geometry, and data) and skills (knowing, applying, and reasoning) following the same assessment framework as the intervention assessments (see section B.1).[15] We leveraged common items across baseline and endline to map both rounds of assessments onto a common scale using a two-parameter logistic item response theory (IRT) model. We standardized all IRT-scaled scores with respect to the control distribution.

3.5 Teachers' estimations of students' achievement

In a prior study, we had elicited teachers' estimations of students' test performance in India and Bangladesh and found that teachers could not accurately predict their students' scores (Djaker et al., 2024). We suspected that teachers' lack of awareness about their students'

[15] Tables C.1 and C.2 show the allocation of items.

performance levels may explain why they do not devote more attention to low achievers, so we included an expanded set of measures in this study. We asked each teacher to estimate the score of 10 randomly sampled students in their classroom on the intervention assessments.[16] We used these estimations to calculate teachers' implied rankings of these students (since they may not know their students' scores, but they may know how students compare) and the implied within-classroom SD in math test scores (since teachers may not know how students compare, but they may have a sense of learning variability in their classrooms).

Given that estimating test scores is challenging, we also asked teachers more concrete questions, such as estimating the share of their students that are proficient (can answer correctly most questions) in each math topic, the share of students who would answer specific items correctly, and whether specific students would answer specific items correctly.

3.6 Instructional time allocation and student-teacher interactions

At endline, trained enumerators conducted in-person classroom observations in all sample schools using two protocols. The first, adapted from the Stallings observation system (Stallings and Mohlman, 1990; Stallings et al., 2014), coded the share of each 40-minute lesson that the teacher spent on instruction, classroom management, or off task, as well as the materials that the teacher was using and the students with whom they were engaging.

The second, developed for this study (Ganimian, 2023), tallied *every* interaction during the lesson—a verbal exchange between the teacher and one student or a small group of students, instead of the whole class—and recorded whether it was started by a teacher or a student; whether it was framed as a statement or question; whether it was short (5 seconds or less), medium (between 6 and 10 seconds), or longer; whether it was uni-directional (one person speaks), bi-directional (one person speaks and the other one responds), or multi-directional (one person speaks, the other one responds, and the first one replies); whether it was academic (focused on the subject being taught), logistic (focused on class-management aspects), or social (focused on teachers' or students' lives); whether it was adversarial (teacher reprimanding a student), supportive (teacher encouraging a student), mixed, or neither; and whether it involved the blackboard, textbooks, notebooks, or other materials. Before each observation, we matched each student in the classroom to their IDs, so that we could understand which students interact more often with their teachers. To our knowledge, ours is the first *direct* measure of whether teachers in LMICs "teach to the top". In a subset of observations, we also

[16] Before this exercise, we asked teachers to indicate which students from a list were in their classroom, which deliberately included fake names, to check whether misestimations were due to teachers not knowing their students' names. During the exercise, we also provided each teacher with a copy of the intervention assessments to minimize the possibility that misestimations were due to lack of familiarity with the instrument.

recorded the topic being covered during the lesson to understand whether teaching to the top also entailed teaching subjects that are too difficult for most students.[17]

This instrument measured student-teacher interactions reliably in two "G(eneralizability) studies" (Cronbach et al., 1972; Shavelson and Webb, 1991; Brennan, 2001), which improve upon traditional reliability analyses by distinguishing between different sources of error variance (e.g., items or raters) and interactions between them (e.g., raters being more stringent on some items). The first, which included all teachers who were observed, indicates that the reliability of the different domains scored (initiator of interactions, direction, content, tone, framing, and length) ranged between 0.179 and 0.481 (Table C.5). The second, for which we added a second rater in a randomly selected subset of classrooms, indicates that the reliability of the same domains ranged from 0.52 to 0.87 for relative judgments (determining which teachers interacted more with their students) and from 0.08 to 0.41 for absolute judgments (determining the specific number and types of interactions for each teacher; Table C.6). A recent review of G-studies of classroom observations in high-income countries found that the reliability of a single observation for relative judgments ranged from 0 to 0.81, with a median of 0.38 (Ganimian et al., 2026), so our instrument compares favorably.

4 Empirical strategy

We estimate the intent-to-treat (ITT) effect of being offered the intervention by fitting:

$$Y_{ij} = \alpha_{r(j)} + X'_{ij}\gamma + T_j\beta + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is the outcome for student i in school j, $\alpha_{r(j)}$ is a randomization-stratum fixed effect, X_{ij} is a vector of baseline covariates selected separately for each outcome using double-selection LASSO (Belloni et al., 2014), T_j is an indicator for school j's assignment to treatment, and ε_{ij} is the error term. The coefficient β captures the ITT effect. We estimate equation (1) by OLS with standard errors clustered at the school level.

To examine how this effect varies across students with different baseline achievement, we divide students into bottom, middle, and top terciles of the within-class baseline math-achievement distribution in the evaluation assessments and estimate one pooled regression

[17] The topics in the observation protocol were matched to the frameworks we used for the intervention (Tables B.1-B.2) and evaluation (Tables C.1-C.2) assessments.

that includes treatment-by-tercile interactions:

$$\begin{aligned} Y_{ij} = \alpha_{r(j)} + X'_{ij}\gamma + T_j\beta_1 + \sum_{k=2}^{3} \delta_k \mathbf{1}[\text{tercile}_{ij} = k] \\ + \sum_{k=2}^{3} \theta_k \big(T_j \times \mathbf{1}[\text{tercile}_{ij} = k]\big) + \varepsilon_{ij}. \end{aligned} \tag{2}$$

The treatment effect for tercile k is recovered as $\beta_1 + \theta_k$ (with $\theta_1 = 0$), and θ_k itself gives the difference in treatment effects between tercile k and the bottom tercile.[18]

5 Results

5.1 Diagnosis

5.1.1 Do teachers know their students' achievement?

Teachers cannot accurately estimate their students' test performance. The correlation between students' test scores on the baseline intervention assessments in math and teachers' estimations of those scores was 0.23, meaning that the latter explained about 5% of the variation in the former (Figure 1, panel A).[19] In fact, teachers seem unable to predict students' within-class rankings: the correlation between students' actual rankings and those implied by teachers' estimations was 0.16 (panel B). Lastly, the actual within-class SDs in students' achievement and the ones implied by teachers' estimations are nearly orthogonal: their correlation is -0.01 (panel C).[20] This pattern, which is consistent with our previous study in India and Bangladesh (Djaker et al., 2024), suggests that teachers may not be catering to the heterogeneity in their students' achievement in part because they are simply unaware of it.

The typical teacher overestimates the test scores of their lowest-achieving students. If we group students by their tercile in the within-class baseline math achievement distribution in the intervention assessments, the average student in the bottom tercile answered 37.1% of the questions correctly, but their teacher believed that they scored 6.9 pp. higher (panel D). This gap is equivalent to 50% of the within-class SD. By contrast, teachers underestimated the scores of middle and top tercile students by 7.9 and 19.7 pp., respectively. These results—

[18]We prefer this pooled specification to subgroup regressions because it estimates the control-variable coefficients on the full sample, yielding tighter standard errors when control effects are approximately uniform.

[19]The baseline was conducted in May, three months after the start of the school year (see section 2.1), so it seems reasonable to expect teachers to be familiar with their students' performance. The length of teachers' exposure to students improves their estimations, but not by much: the correlation between students' scores on the endline evaluation assessments and teachers' estimations in the control group was 0.33.

[20]The difficulty of estimating students' scores does not explain these results. More concrete elicitation questions show a similar pattern of limited accuracy: teachers often overestimated topic-level proficiency and specific student-item performance, though not every item-level question (Table A.5).

consistent with our earlier work (Djaker et al., 2024) and replication studies in Benin, Nigeria, and The Gambia (Rodriguez-Segura, 2024), shown in Figure A.2—suggest that teachers may devote insufficient attention to low achievers in their classroom because they do not know how many of them there are or the extent to which they underperform.

5.1.2 (How) do teachers interact with students?

The typical teacher rarely interacts with their students during class. In the average 40-minute math lesson in the control group at endline,[21] the teacher had 8.5 verbal exchanges with a student or a small group of students (Figure 2, panel A). This suggests teachers have few opportunities to update their estimations of their students' achievements during class.

Teachers vary widely in the number of interactions with their students. The teacher with the fewest interactions had only 1 and the one with the most interactions had 32; the SD across teachers was 5.3 interactions. Yet, teacher characteristics explain little of this variation. If we group teachers by whether they possess a characteristic (if discrete) or whether they are above the median (if continuous), most have a small and statistically insignificant association with interaction counts, though older and more experienced teachers interact somewhat more often, while teachers with education majors and higher self-efficacy interact less often (Table A.8).

Most students do not interact with their teacher during class. The average control student had 0.24 verbal exchanges with their teacher and 79% had none at all (Table A.7, panel A).[22] Yet, their achievement was the strongest predictor of whether they interacted and how often they did so (Table A.9). The average student in the bottom tercile of the within-class baseline math achievement distribution in the evaluation assessments had less than half the interactions of their top-tercile counterpart, and the former was 1.3 times less likely to interact with their teacher at all than the latter (Figure 2, panel C and Table A.7, panel A). Thus, teachers had very few chances to update their achievement estimations for this group.

The differences in interactions by students' achievement are driven by teachers calling on low achievers less frequently. Control teachers started 69% of the interactions that we observed, and the modal interaction involved a teacher asking a student a question about academics and the student responding (Figure 2, panel B and Table A.7, panels B-E). The modal interaction lasted about 10 seconds (panel F), which suggests that teachers mostly asked yes-no or low-inference questions. Together, these results offer some of the most direct evidence to date that teachers in LMICs teach to the top of the achievement distribution.

[21]We did not measure interactions at baseline, so we use the endline control group to characterize ordinary classroom instruction during the endline data-collection round.

[22]This finding is consistent with the results from our observations of teachers' time allocation, which indicate that 69% of lesson time is devoted to whole-group instruction (Table A.6, panel C).

5.2 Impact

5.2.1 Did the intervention work as expected?

Teachers responded to the information in the reports as intended. In treatment schools—the only ones that received intervention assessments and reports—the content and students that were identified as in need of support in the baseline intervention reports improved more from baseline to midline (the last time that the intervention assessments were administered) than the rest. This is consistent with the reports successfully focusing teachers' attention.

Students improved more on the topics that the reports flagged for low performance. As shown in Figure B.2, each report indicated the topic in which the class performed least well (e.g., measurements and geometry). Treatment students improved more in the intervention assessments from baseline to midline on these flagged topics than on non-flagged topics: 25 pp. versus 9.8 pp.—a difference-in-difference (DID) estimate of 14.5 pp. (Table 1, panel A).

The same pattern emerges for sub-topics and items. As Figure B.2 shows, reports identified the sub-topics on which students needed most help (e.g., applying geometric properties to find a missing angle). Treatment students improved on these flagged sub-topics by 37.6 pp., while non-flagged ones improved by 10.8 pp.—a DID effect of 19.4 pp. (panel A). Even the *color* that reports used to flag items seems to have had an effect. Reports showed items with low performance (less than one-third of students answering correctly) in red, items with medium performance (one-to-two-thirds) in yellow, and items with high performance (more than two-thirds) in green (Figure B.2). Red and yellow items improved by 34.7 pp. more than green items, and red items improved by 28.1 pp. more than yellow items (panel A).[23]

Students flagged in the reports also improved more than their peers. Figure B.3 shows that reports identified the students most in need of support in each topic. Treatment students flagged in at least one topic improved by 22.5 pp. from baseline to midline, compared with 12.7 pp. among students not flagged in any topic—a DID estimate of 9.9 pp. (Table 1, panel B). The gains were larger for students flagged in more topics: students flagged in one, two, and three topics improved by 21.2, 25.5, and 33.3 pp., respectively. The corresponding DID estimates were 8.5, 12.8, and 20.6 pp., with the last estimate based on only 27 students.

Further, flagged students improved most in the specific topics for which they were flagged. Among treatment students who were flagged in at least one topic, performance rose by 41.8 pp. in the flagged topics, compared with 9.8 pp. in the topics in which they were not flagged—a DID estimate of 32.9 pp. (panel D, flagged-topic rows). Because flagging reflected baseline performance, this comparison remains descriptive. Still, the pattern is consistent with reports focusing teacher attention on topics and students flagged for support.

[23]Because item colors were assigned mechanically from baseline class performance, red and yellow items began with lower scores than green items and therefore had more room to improve. These comparisons may therefore partly reflect regression to the mean or ceiling effects rather than teachers responding to the colors.

5.2.2 Did the intervention raise math achievement?

The intervention had broad-based effects on math achievement. The average student improved by 0.23 SDs in math and we cannot rule out the possibility that students in the bottom third of the within-class baseline math achievement distribution in the evaluation assessments improved by as much as those in the other two thirds (Table 2, panel A). Non-parametric plots of average treatment effects confirm this pattern, with treatment students outperforming their control peers by a similar margin across the baseline distribution (Figure A.4, panel A). Quantile treatment effects present a similar pattern, but with a smaller contrast (panel B).

The intervention improved achievement across all assessed math topics and cognitive skills. Effects were 0.24 SDs on number, 0.16 SDs on measurement and geometry, and 0.16 SDs on data (Table 2, panel B). They were 0.18 SDs on knowing, 0.19 SDs on applying, and 0.26 SDs on reasoning (panel C). Although the point estimates were largest on number and reasoning, we cannot reject the null hypotheses that effects were equal across topics ($p = 0.131$) or across skills ($p = 0.144$). The gains were therefore broad-based.

Consistent with section 5.2.1, flagged students fared better than their control counterparts. Comparing observed treatment flags with evaluation-assessment flags among controls or applying evaluation-assessment flags in both groups yields similar effects: 0.21 and 0.24 SDs, respectively (Table A.11). Non-flagged treatment students also outperformed their control counterparts by similar margins; we cannot reject equality under either specification. Thus, whatever benefited targeted students also benefited non-targeted peers.

5.2.3 Did the intervention have spillover effects on reading?

The intervention also had positive spillover effects on reading. The average student improved by 0.17 SDs in reading (Table 3, panel A). Effects are statistically significant only for the middle and top terciles of the within-class baseline math achievement distribution in the evaluation assessments, but we cannot reject the null hypotheses that students in the bottom tercile improved by as much as their counterparts in the other two terciles.

Gains were concentrated on the inputs and skills that math word problems exercise. Grade-6 teachers in Bangladesh are subject-specific, so the intervention is unlikely to have affected reading performance by changing the way reading was taught. Therefore, we check whether it improved achievement on the topics and skills that might travel across subjects. That is what we find: effects were 0.15 SDs on word items and 0.13 SDs on sentence items (panel B) and 0.15 SDs on vocabulary and 0.14 SDs on comprehension items (panel C). These results suggest that, through changes in the instructional practices of math teachers (e.g., increased use of word problems), students learned inputs and skills that improved their reading proficiency.[24]

5.2.4 Did the intervention raise student effort and motivation?

We find some evidence that the intervention increased students' effort and motivation, particularly among lower-achieving students. School registers indicated that students attended 70% of days at both midline and endline; enumerators found 62% of enrolled students present on the survey day at midline and 61% at endline. The intervention did not affect either attendance measure on average at either round; in fact, we can rule out increases above 4 pp. in register-based attendance and 5.0 pp. in observed attendance. At endline, however, it increased register-based attendance among low-achieving students by 3.5 pp. (Table A.15).[25]

The intervention also improved social-emotional skills. It increased self-efficacy (confidence in completing tasks) by 0.14 SDs, growth mindset (belief that ability and intelligence can develop) by 0.11 SDs, and task-goal orientation (desire to learn and master class material) by 0.10 SDs (Table 4). Effects were null for self-management (preparedness, attention, and self-control), fear of failure (worry and self-doubt when failing), motivation to master tasks (hard work and persistence), and grit (long-term perseverance), although we cannot rule out positive effects of 0.10 SDs for any of them. Among low-achievers, the respective increases were 0.14, 0.20, and 0.13 SDs; only growth mindset differed significantly from high-achievers.

5.3 Mechanisms

5.3.1 Did teachers update their estimations of student achievement?

Teacher estimates did not become more accurate. The intervention did not impact absolute gaps between estimated and actual scores or ranks (Table 5, panel A), or the gap between estimated and actual within-class score SDs (panel C). This is perhaps unsurprising because reports identified topics and students in greatest need of support. They did not disclose individual scores or the within-class distribution needed to update teachers' beliefs about students' absolute performance or relative standing within their class at the endline elicitation.

Instead, the intervention shifted teachers' estimations downward across the achievement distribution. Estimated scores shifted downward relative to actual scores by 0.47 within-class SDs, while estimated within-class percentile ranks shifted downward relative to actual ranks by 0.9 pp. (Table 5, panel B). The score shift was similar across achievement terciles. Together

[24]Reading gains were also similar for students flagged for support in math and non-flagged students under both approaches used to define comparable flagged groups (Table A.12). Because the intervention assessments covered only math, these comparisons use the math-based flags defined in section 3.4.

[25]We also checked whether the intervention impacted the effort that students exerted on the endline assessments, but found null results, mostly because nearly all students attempted nearly all items.

with the positive impacts across all three achievement terciles, and across flagged and non-flagged students on the evaluation assessments, the absence of differential updating suggests that whatever teachers did to improve low achievers also benefited their higher-achieving peers.

5.3.2 Did the intervention change student-teacher interactions?

Consistent with the downward shifts in teachers' estimations across all achievement terciles, the intervention led them to distribute interactions more evenly across students. It left average interactions per student per lesson unchanged (Table 6, panel A), but interactions rose by 0.01 for low-achievers and fell by 0.02 and 0.07 for medium- and high-achievers, with a significant low-high difference. These count changes imply that low-achievers' share rose by roughly as much (from 22% to 27%) as high-achievers' share fell (from 45% to 40%).

5.3.3 Did the intervention change the topics teachers covered?

The intervention seems to have led teachers to focus on more foundational content. In a random subsample, observers recorded whether teachers covered number, measurement and geometry, or data. The intervention reduced data coverage by 15.3 pp. from 18.2%, left measurement and geometry unchanged, and increased number coverage by 13.6 pp. from 72.7% (Table A.14). The latter effect was not statistically significant, but the small observation subsample means that the estimate is imprecise and difficult to distinguish from zero.

Together, the three mechanisms suggest that the intervention changed how teachers taught the full class rather than only inducing additional support for low-achievers. Teachers revised their estimations downward across the achievement distribution, distributed interactions more evenly, and focused on more foundational content. This interpretation is consistent with positive impacts for all achievement terciles and for both flagged and non-flagged students.

6 Discussion

6.1 Cost-effectiveness

The formative-assessment intervention was highly cost-effective. Following the ingredients method in Dhaliwal et al. (2013), which identifies and values each input required to deliver the intervention, we estimate that two rounds of assessment, reporting, and teacher orientation cost USD 5.06 per student.[26] This implies a cost-effectiveness estimate of 4.54 SDs in math and 3.36 SDs in reading per USD 100 spent. These estimates rank at the 86th and

[26]The total cost was BDT 1.87 million, or USD 18,155 in 2023 dollars, for approximately 3,588 students. Costs exclude evaluation activities, teachers' opportunity costs, and existing school-system inputs that would have been incurred as part of routine schooling.

80th percentiles in the distribution of cost-effectiveness estimates from rigorously evaluated education interventions in LMICs (Angrist et al., 2025). The gains remain large in learning-adjusted years of schooling (LAYS), using 0.8 SDs of learning as the benchmark for one year of high-quality schooling (Angrist et al., 2020). The effects equal 5.68 LAYS in math and 4.20 in reading per USD 100—more than 11 and eight times the benchmark median of 0.49 LAYS.

No comprehensive database tracks the cost-effectiveness of social-emotional interventions in LMICs, so we benchmark our results against others in economics. The intervention generated 2.67 SDs in self-efficacy, 2.17 in growth mindset, and 1.98 in task-goal orientation per USD 100 spent. These estimates exceed the 1.36-1.59 SDs from home visits through Bangladesh's nutrition system (Bos et al., 2024) and are similar to the 3.29-3.41 SDs from a perspective-taking curriculum delivered to primary-school students in Turkey (Alan et al., 2021).[27]

6.2 Scale-up

It is fiscally feasible for Bangladesh to scale the formative-assessment intervention nationwide. In 2023, it enrolled 2.10 million grade-6 students in general-secondary schools (BANBEIS, 2023). If we multiply this figure by our per-student intervention cost of BDT 521, extending the intervention to every sixth grader would cost BDT 1.10 billion (USD 10.6 million), which amounts to 0.23% of the government's allocation for secondary and higher education for the 2025-2026 school year (Ministry of Finance, 2025). Yet, this projection overstates recurring costs given the economies of scale that centralized procurement and reporting would generate.

A nationwide scale-up could leverage Bangladesh's administrative structure, but it may require a sizable temporary field workforce. In our study, 26 enumerators and 6 supervisors delivered two assessment rounds across 156 treatment schools, with each person deployed for 17 days. Applying these staffing ratios to the country's 20,448 general secondary institutions implies approximately 3,400 enumerators and 800 supervisors, or 71,000 staff-days per year (BANBEIS, 2023). The country's 480 *upazila* secondary education offices, together with its district and zonal offices, provide an existing structure through which to coordinate this effort; each *upazila* office would oversee 43 schools on average (DSHE, 2026). Yet, we do not know whether these offices have sufficient personnel to administer and supervise the assessments. A scale-up of the evaluated model may therefore require hiring external assessors.

Bangladesh is beginning to incorporate formative assessment and remediation into its secondary-education system, but this initiative differs from the intervention that we evaluated. The World Bank-supported program, approved in December 2025, seeks to train teachers in grades 6 and 8 to administer formative assessments in math and Bangla and use test results to deliver remedial education. It plans to reach 1,200 schools in its first year (by March 2026),

[27] Other studies have found impacts on social-emotional skills (Alan and Ertac, 2018; Alan et al., 2019; Özler et al., 2018; Premand and Barry, 2022), but they do not report program costs.

3,000 schools in its second year, and 7,200 schools in its third year (World Bank, 2023, 2025). We expect this rollout to shed light on whether asking teachers to assess students, analyze results, and adjust their instruction can have similar results to the approach that we evaluated. Prior research from India points to two main challenges: ensuring that the tests administered by teachers are as reliable as those conducted by external assessors (Singh and Berg, 2024) and that teachers use test results to inform their instructional practices (Duflo et al., 2015).

6.3 Diagnostic information

Our study meaningfully advances global evidence on the impact of diagnostic (low stakes) information in LMICs.[28] It leveraged common elements from prior interventions in Argentina (de Hoyos et al., 2021; de Hoyos et al., 2023) and India (Banerjee et al., 2018) and showed that interventions that are brief (reports were five pages), user-friendly (results were presented in words, circumventing challenges that teachers may face interpreting graphs), concrete (reports highlighted the topics and students in greatest need of support), and actionable (answer keys explained both correct and incorrect answers) can raise academic and social-emotional skills. It verified that teachers responded to the specific information included in the interventions, by tracking how much students improved on aspects highlighted in the reports relative to others. And it identified three mechanisms of impact: teachers' awareness of their students' learning, their interactions with students, and the level at which they pitch classroom instruction.

Remarkably, our study demonstrates that information provision can impact instruction and learning even in the absence of changes in the institutional incentives that teachers face. Together with encouraging evidence of similar interventions (de Hoyos et al., 2021; de Hoyos et al., 2023), it suggests that earlier null effects (Muralidharan and Sundararaman, 2010) cannot be attributed solely to the lack of incentives tied to information, as previously believed. Further, it is less politically and administratively disruptive than efforts to change incentives, such as curricular reforms (Rodriguez-Segura and Mbiti, 2022), tracking (Duflo et al., 2011), merit pay (Muralidharan and Sundararaman, 2011), or contract teachers (Duflo et al., 2015). In fact, most LMICs already collect the data needed for diagnostic-information interventions. By 2024, 103 of 131 such countries had reported administering a nationally representative assessment of reading and math at the end of primary or secondary school (UNESCO, 2026).

[28] We see these efforts, in which principals or teachers are provided with granular and actionable data (e.g., topic-, skill- or item-level results) to guide their practice, as testing a different theory of change from those disclosing school/student test scores to promote accountability through parental demand and competition among schools (Banerjee et al., 2010; Mizala and Urquiola, 2013; Andrabi et al., 2017; Camargo et al., 2018).

6.4 Teaching to the top

To our knowledge, our study also provides the first direct evidence that teachers in LMICs "teach to the top" of the ability distribution. It documents that teachers not only interact with high achievers more often than with low achievers, but that this unequal distribution of interactions is driven by interactions that teachers start (not by those started by students). The strongest evidence to date of this phenomenon stemmed from an experiment in Kenya that found achievement patterns consistent with teachers pitching instruction to the top of heterogeneous classes but did not observe their instructional decisions (Duflo et al., 2011).[29] We complement this evidence with measures of teachers' beliefs and classroom practices.

We also show that the tendency to teach to the top is more malleable than many thought. The intervention that we evaluated impacted both how teachers allocated interactions across low-, medium-, and high-achieving students and the (difficulty of the) topics that they taught. This is striking, given that these behaviors have been attributed to teachers seeking to avoid being held accountable for not completing the curriculum or having too few of their students pass high-stakes exams (Glewwe et al., 2009; Kremer et al., 2009; Pritchett and Beatty, 2015). Our results should prompt us to widen our consideration of the incentives that teachers face. For example, it seems entirely plausible that teachers in our setting see the intervention as a change in the priorities of the government and are merely responding to this perceived shift.

6.5 Differentiated instruction

Finally, our results raise the possibility that a non-trivial share of the effect of differentiated instruction may be due to teachers updating their priors about their students' learning levels. As we understand it, differentiation entails assessing students, grouping them based on their test performance, and assigning each group to activities that match their level of preparation. Together with other approaches seeking to improve the fit of instruction to students' needs—collectively known as "Teaching at the Right Level" (TaRL)—it is regarded as one of the world's most effective education interventions (Akyeampong et al., 2023; Angrist and Meager, 2023). In our study, the first of these steps (assessing students and reporting test results to teachers) led teachers to downwardly adjust their estimations of students' test scores, interact with students more equitably, and teach more foundational topics, raising learning for all students. These results invite us to consider the possibility that teachers' awareness of their students' achievement may account for a nontrivial share of the impact of differentiated instruction.

[29] The authors argue that two patterns are inconsistent with teachers pitching instruction to the median student and consistent with their focusing on those at the top: tracking raised achievement across the baseline distribution, including in the lower track; and students near the median performed similarly in both tracks despite large differences in peer achievement. They also find larger gains in basic skills among lower achievers and in advanced skills among higher achievers, consistent with teachers adjusting instruction to each track.

Mounting evidence suggests that teachers in LMICs are unaware of students' achievement. In prior work in India and Bangladesh (Djaker et al., 2024), we were among the first to show that teachers in LMICs are far less likely to correctly estimate their students' achievement than their peers in HICs (Hoge and Coladarci, 1989; Südkamp et al., 2012; Kaufmann, 2024).[30] That study sparked replications in Benin, Nigeria, and The Gambia (Rodriguez-Segura, 2024), demonstrating that the patterns we observed in South Asia extended to Sub-Saharan Africa. In the present study, we show that these estimations are responsive to information on students' achievement and that the uniform downward adjustment in estimations coincided with a more equal allocation of student-teacher interactions and increased coverage of foundational topics. This pattern echoes that of a recent study in India that found that teachers overestimated their students' achievement, but that after implementing a remedial-education program—that entailed assessing their students—they adjusted their estimates downward (Beg et al., 2024). These two studies raise the question of whether the effects of differentiation may be driven by teachers updating their priors about students' skills and adjusting instruction accordingly.

This question is crucial because prior impact evaluations of differentiated instruction have found that teachers struggle to implement one of its key components (grouping students by test performance) during the school day unless they are provided with dedicated or extra time. In Bihar, India, just 1.6% of teachers grouped students as intended when assigned to materials and 3.8% did so when they also received training. In Uttarakhand, only 10% of teachers with training and materials grouped students and 5.1% did so when also granted volunteer support. By contrast, in Haryana 91% of teachers grouped students when given a dedicated hour for differentiation, compared to merely 2% of them outside that hour (Banerjee et al., 2017).[31] Given the impacts of testing students and reporting results to teachers, and the challenges of grouping students by test performance, our study suggests that a pared-down version of differentiation could increase teacher take-up, cost-effectiveness, and potential for scale-up.

7 Conclusion

Teachers cannot align instruction with students' needs when they lack actionable information about what their classes understand. We provide experimental evidence that two rounds of formative assessments, reports, and instructional guidance helped teachers in Bangladesh respond to this constraint. The intervention raised math by 0.23 SDs and reading by 0.17 SDs,

[30] Wadmare et al. (2022) had found that 40% of teachers wrongly believed that their low-achieving students attained foundational literacy. Yet, their sample precluded comparisons of errors across achievement groups, their focus on foundational skills may have understated gaps relative to curricular expectations, and their use of performance levels yielded a coarse measure of correspondence between actual and estimated achievement.

[31] Even dedicated time may fail to ensure implementation if teachers see it as competing with their other responsibilities. In Ghana, just 6% of teachers grouped students even if an hour was meant to be devoted to it. Teachers reported that they were still expected to cover the grade-level curriculum (Duflo et al., 2024).

improved several social-emotional skills, and generated similar gains throughout the baseline achievement distribution, including among students who were not identified as needing support in the reports and other classmates in the same classrooms throughout the school year.

The intervention appears to have worked by redirecting teachers' attention. Teachers adjusted their estimations of their students' performance downward across the board, leading them to distribute interactions more equitably and to review foundational content. The broad gains indicate that focusing on unmet learning needs can help lower-achieving students without reducing gains among classmates elsewhere in the baseline achievement distribution.

Overall, actionable assessment results paired with concrete guidance offer a promising way to help teachers serve academically diverse classrooms. At USD 5.06 per student, nationwide delivery appears fiscally feasible, but replicating the evaluated model would require reliable independent assessment and substantial implementation capacity. Further research should test whether school systems can preserve both under routine delivery at scale without sacrificing assessment reliability or the instructional usefulness of reports to teachers over time.

References

Abadzi, H. (2009). Instructional time loss in developing countries: Concepts, measurement, and implications. *The World Bank Research Observer 24*(2), 267–290.

Aghion, P. and J. Tirole (1997). Formal and real authority in organizations. *Journal of Political Economy 105*(1), 1–29.

Akyeampong, T., T. Andrabi, A. V. Banerjee, R. Banerji, S. Dynarksi, R. Glennerster, S. Grantham-McGregor, K. Muralidharan, B. Piper, S. Ruto, J. Saavedra, S. Schmelkes, and H. Yoshikawa (2023). *2023 cost-effective approaches to improve global learning. What does recent evidence tell us are "smart buys" for improving learning in low- and middle-income countries?* London, UK; Washington, DC; New York, NY: Foreign, Commonwealth Development Office (FCDO), World Bank, United Nations International Children's Emergency Fund (UNICEF), United States Agency for International Development (USAID). London, UK; Washington, DC; New York, NY: Foreign, Commonwealth & Development Office (FCDO), World Bank, United Nations International Children's Emergency Fund (UNICEF), United States Agency for International Development (USAID).

Alan, S., C. Baysan, M. Gumren, and E. Kubilay (2021). Building social cohesion in ethnically mixed schools: An intervention on perspective taking. *The Quarterly Journal of Economics 136*(4), 2147–2194.

Alan, S., T. Boneva, and S. Ertac (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics 134*(3), 1121–1162.

Alan, S. and S. Ertac (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy 126*(5), 1865–1911.

Andrabi, T., J. Baron, I. Macdonald, and Z. Qureshi (2025). Free to choose: Introducing technology for foundational learning in Pakistan. Working paper, April 13.

Andrabi, T., J. Das, and A. I. Khwaja (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review 107*(6), 1535–1563.

Angrist, N., C. Cullen, and J. Magat (2025). Cheaper (and more effective) by the dozen: Evidence from 12 randomized a/b tests optimizing tutoring for scale. edworkingpaper no. 25-1333. *Annenberg Institute for School Reform at Brown University*.

Angrist, N., S. Djankov, P. K. Goldberg, and H. A. Patrinos (2021). Measuring human capital using global learning data. *Nature 592*(7854), 403–408.

Angrist, N., D. K. Evans, D. Filmer, R. Glennerster, F. H. Rogers, and S. Sabarwal (2020). How to improve education outcomes most efficiently? A comparison of 150 interventions using the new learning-adjusted years of schooling metric. (Policy Research Working Paper No. 9450). Washington, DC: The World Bank.

Angrist, N., D. K. Evans, D. Filmer, R. Glennerster, H. Rogers, and S. Sabarwal (2025). How to improve education outcomes most efficiently? a review of the evidence using a unified metric. *Journal of Development Economics 172*, 103382.

Angrist, N. and R. Meager (2023). Implementation matters: Generalizing treatment effects in education.

BANBEIS (2023). Bangladesh education statistics 2023. Dhaka, Bangladesh: Bangladesh Bureau of Educational Information and Statistics (BANBEIS). Retrieved from `https://banbeis.portal.gov.bd/sites/default/files/files/banbeis.portal.gov.bd/npfblock/Bangladesh%20Education%20Statistics%202023%20%281%29.pdf`.

Banerjee, A. V., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017). From proof to concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives 31*(4), 73–102.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in india. *American Economic Journal: Economic Policy 2*(1-30), 1–30.

Banerjee, A. V., S. Cole, E. Duflo, and L. L. Linden (2007). Remedying education: Evidence from two randomized experiments in india. *The Quarterly Journal of Economics 122*(3), 1235–1264.

Banerjee, A. V. and E. Duflo (2011). Top of the class. *Poor economics: A radical rethinking of the way to fight global poverty*. Philadelphia, PA: Public Affairs.

Banerjee, A. V., A. J. Ganimian, and S. Mukerji (2018). The impact of diagnostic feedback on differentiated instruction: Experimental evidence from India. *Unpublished manuscript*. New Delhi, India: Abdul Latif Jameel Poverty Action Lab (J-PAL) South Asia.

Barro, R. J. and J. W. Lee (2026). Barro-Lee educational attainment indicators. World Bank Education Statistics, Barro-Lee educational attainment indicators. Retrieved from `https://databank.worldbank.org/source/education-statistics-%5E-all-indicators`. Retrieved: May 21, 2026.

Beg, S. A., A. E. Fitzpatrick, J. T. Kerwin, A. Lucas, and K. W. Rahman (2024, December). When given discretion teachers did not shirk: Evidence from remedial education in secondary schools. Working Paper 33242, National Bureau of Economic Research.

Beg, S. A., A. E. Fitzpatrick, and A. M. Lucas (2024). Managing to learn.

Beg, S. A., A. M. Lucas, W. Halim, and U. Saif (2022). Engaging teachers with technology increased achievement, bypassing teachers did not. *American Economic Journal: Economic Policy 14*(2), 61–90.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Bergman, P. (2020). Parent-child information frictions and human capital investment: Evidence from a field experiment. *Journal of Political Economy*.

Bhattacharjea, S., W. Wadhwa, and R. Banerji (2011). Inside primary schools: A study of teaching and learning in rural india.

Bold, T., D. Filmer, G. Martin, E. Molina, B. Stacy, C. Rockmore, J. Svensson, and W. Wane (2017). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in africa. *Journal of Economic Perspectives 31*(4), 185–204.

Bos, J. M., A. S. Shonchoy, S. Ravindran, and A. Khan (2024). Early childhood human capital formation at scale. *Journal of Public Economics 231*, 105046.

BRAC (2015). Education watch 2014: Whither grade V examination? An assessment of primary education completion examination in Bangladesh. Dhaka, Bangladesh: Building Resources Across Communities (BRAC).

Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: New York: Springer-Verlag.

Bruns, B. and J. Luque (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. Washington, DC: Washington, DC: The World Bank.

Büchel, K., M. Jakob, C. Kühnhanss, D. Steffen, and A. Brunetti (2020). The relative effectiveness of teachers and learning software: Evidence from a field experiment in El Salvador. (UBSS Working Paper No. 36). Bern, Switzerland: University of Bern.

Camargo, B., R. Camelo, S. Firpo, and V. Ponczek (2018). Information, market incentives, and student performance: Evidence from a regression discontinuity design in Brazil. *The Journal of Human Resources 53*(2), 414–444.

Cilliers, J. and J. Habyarimana (2026). Addressing weak links in government implementation at scale: Experimental evidence from a school governance reform in Tanzania. *Journal of Development Economics 182*, 103824.

Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnam (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

de Barros, A. and T. Lubozha (2026). Targeting foundational skills at scale: Skill specificity and transfer. CEGA Working Paper No. 259. Berkeley, CA: Center for Effective Global Action (CEGA).

de Hoyos, R., S. Djaker, A. J. Ganimian, and P. A. Holland (2023). Can performance-management tools and training complement diagnostic feedback? experimental evidence from public schools in argentina. *Economics of Education Review 9*, 102518.

de Hoyos, R., A. J. Ganimian, and P. A. Holland (2021). Teaching *with* the test: Experimental evidence on diagnostic feedback and capacity-building for schools in Argentina. *World Bank Economic Research 35*(2), 499–520.

Dhaliwal, I., E. Duflo, R. Glennerster, and C. Tulloch (2013). Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education. *Education Policy in Developing Countries*, 285–338.

Dizon-Ross, R. (2019). Parents' beliefs and children's education: Experimental evidence from Malawi. *American Economic Review 109*(8), 2728–2765.

Djaker, S. (2023). *Targeting teachers' perceptions of student ability to improve learning levels in low- and middle-income countries*. Ph. D. thesis, New York University.

Djaker, S., A. J. Ganimian, and S. Sabarwal (2024). Out of sight, out of mind? the gap between students' test performance and teachers' estimations in india and bangladesh. *Economics of Education Review 102*, 102575.

DSHE (2019). Frameworks for National Assessment of Secondary Students 2019. Subject: Bangla, English, and Mathematics. Grades: 6, 8, and 10. Dhaka, Bangladesh: Monitoring and Evaluation Wing, Directorate of Secondary and Higher Education (DSHE).

DSHE (2026). History and administrative structure. Dhaka, Bangladesh: Directorate of Secondary and Higher Education (DSHE). Retrieved from `https://dshe.gov.bd/pages/static-pages/691997c4933eb65569ddef95`. Retrieved: August 4, 2026.

Duflo, A., J. Kiessel, and A. M. Lucas (2024). Experimental evidence on alternative policies to increase learning at scale. *The Economic Journal 134*(661), 1985–2008.

Duflo, E., J. Berry, S. Mukerji, and M. Shotland (2015). A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India. (Impact Evaluation Report No. 22). International Initiative for Impact Evaluation (3ie). New Delhi, India.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review 101*(5), 1739–1774.

Duflo, E., P. Dupas, and M. Kremer (2015). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from kenyan primary schools. *Journal of Public Economics 123*, 92–110.

Evans, D. K. and A. M. Acosta (2026). Changing perceptions of educational returns in low-and middle-income countries: A review of the evidence. *International Journal of Educational Development 122*, 103563.

Ganimian, A. J. (2023). Do teachers in developing countries "teach to the top"? A new approach to measure student-teacher interactions. Spencer Foundation/National Academy of Education Fall Fellows Retreat.

Ganimian, A. J., A. D. Ho, and A. Campos Quintero (2026). The reliability of classroom observations and student surveys in non-research settings: Evidence from a middle-income country. (EdWorkingPaper No. 26-1442). Providence, RI: Annenberg Institute at Brown University. DOI: 10.26300/c9zp-p565.

Ganimian, A. J., I. M. Mbiti, and A. Mishra (2026). Leveraging complementarities between teachers' content knowledge and pedagogical supports: Experimental evidence from India. *Unpublished manuscript*. New York, NY: New York University (NYU).

GEEAP (2020). Cost-effective approaches to improving global learning: What does recent evidence tell us are "Smart Buys" for improving learning in low- and middle-income countries? Washington, DC: World Bank; Foreign, Commonwealth & Development Office; Building Evidence in Education.

Gilligan, D. O., N. Karachiwalla, I. Kasirye, A. M. Lucas, and D. Neal (2022). Educator incentives and educational triage in rural primary schools. *Journal of Human Resources 57*(1), 79–111.

Glewwe, P., N. Ilias, and M. Kremer (2010). Teacher incentives. *American Economic Journal: Applied Economics 2*, 205–227.

Glewwe, P., M. Kremer, and S. Moulin (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics 1*(1), 112–135.

Hoge, R. D. and T. Coladarci (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research 59*(3), 297–313.

IEA (2021). TIMSS 2023: Assessment frameworks. Edited by Mullis, I. V. S., Martin, M. O. & von Davier, M. TIMSS & PIRLS International Study Center. Lynch School of Education and Human Development, Boston College & International Association for the Evaluation of Educational Achievement (IEA).

Kaufmann, E. (2024). Teachers' judgment accuracy: A replication check by psychometric meta-analysis. *PLOS ONE 19*(7), e0307594.

Kremer, M., C. Brannen, and R. Glennerster (2013). The challenge of education and learning in the developing world. *Science 240*, 297–300.

Kremer, M., E. Miguel, and R. Thornton (2009). Incentives to learn. *The Review of Economics and Statistics 91*(3), 437–456.

Ministry of Finance (2025). National budget speech 2025-26. Dhaka, Bangladesh: Government of the People's Republic of Bangladesh. Retrieved from `https://nbr.gov.bd/uploads/budget/English.pdf`.

Mizala, A. and M. Urquiola (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics 103*, 313–335.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review 109*(4), 1–35.

Muralidharan, K. and V. Sundararaman (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from india. *The Economic Journal 120*(F187-F203), F187–F203.

Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy 119*(1), 39–77.

Muralidharan, K. and Y. Zieleniak (2014). Chasing the syllabus: Measuring learning trajectories in developing countries with longitudinal data and item response theory. *Unpublished manuscript*. University of California, San Diego. San Diego, CA.

NCTB (2013). Mathematics: Class six. Dhaka, Bangladesh: National Curriculum and Texbook Board (NCTB).

NCTB (2021). National curriculum framework 2021. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB).

NCTB (2023). National curriculum framework 2021: Pre-primary to grade 12. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB). Retrieved from `https://shed.portal.gov.bd/site/page/f33f6b5b-b650-4c9b-81ba-270ae88fa6e0/National-Curriculum-Framework-2021-%28English-Version%29`.

Neilson, C., F. Gallego, and O. Molina (2015). The impact of information provision on human capital accumulation and child labor in peru. *Innovations for Poverty Action. https://www.poverty-action. org/printpdf/21321*.

Nourani, V., N. Ashraf, and A. Banerjee (2025). Learning to teach by learning to learn. Working paper, February.

Özler, B., L. C. H. Fernald, P. Kariger, C. McConnell, M. Neuman, and E. Fraga (2018). Combining pre-school teacher training with parenting education: A cluster-randomized controlled trial. *Journal of Development Economics 133*, 448–467.

Piza, C., A. Zwager, M. Ruzzante, R. Dantas, and A. Loureiro (2024). Teacher-led innovations to improve education outcomes: Experimental evidence from Brazil. *Journal of Public Economics 234*, 105123.

Premand, P. and O. Barry (2022). Behavioral change promotion, cash transfers and early childhood development: Experimental evidence from a government program in a low-income setting. *Journal of Development Economics 158*, 102921.

Pritchett, L. and A. Beatty (2015). Slow down, you're going too fast: Matching curricula to student skill levels. *International Journal of Educational Development 40*, 276–288.

Rodriguez-Segura, D. (2024). NewGlobe policy reports. NewGlobe policy reports.

Rodriguez-Segura, D. and I. Mbiti (2022). Back to the basics: Curriculum reform and student learning in tanzania.

Sabarwal, S., M. Abu-Jawdeh, and R. Kapoor (2022). Teacher beliefs: Why they matter and what they are. *The World Bank Research Observer 37*(1), 73–106.

Sankar, D. and T. Linden (2014). How much and what kind of teaching is there in elementary education in india? evidence from three states. Washington, DC: The World Bank.

Shavelson, R. J. and N. M. Webb (1991). *Generalizability theory: A primer*. Sage.

Singh, A. and P. Berg (2024). Myths of official measurement: Limits to test-based education reforms with weak governance. *Journal of Public Economics 239*, 105246.

Stallings, J. A., S. L. Knight, and D. Markham (2014). Using the stallings observation system to investigate time on task in four countries. *Unpublished manuscript*. Washington, DC: The World Bank.

Stallings, J. A. and G. G. Mohlman (1990). Issues in qualitative education research: Observation techniques. In H. J. Walberg & G. D. Haertel (Eds.) *The international encyclopedia of educational evaluation* (pp. 639-644). New York, NY: Pergamon Press.

Südkamp, A., J. Kaiser, and J. Möller (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology 104*(3), 743.

Tatto, M. T., R. Peck, J. Schwille, K. Bankov, S. L. Senk, M. Rodriguez, L. Ingvarson, M. Reckase, and G. Rowley (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)*. Amsterdam, the Netherlands: Amsterdam, the Netherlands: International Evaluation Association (IEA).

UIS (2026). UIS Data API. UNESCO Institute for Statistics (UIS) Data Browser. `https://api.uis.unesco.org/api/public/data/indicators`. Retrieved: May 21, 2026.

UNESCO (2023). Efa global monitoring report 2023: Technology in education: A tool on whose terms?

UNESCO (2026). Sustainable Development Goal 4 data. Paris, France: UNESCO Institute for Statistics. February 2026 data release.

Wadmare, P., M. Nanda, R. Sabates, N. Sunder, and W. Wadhwa (2022). Understanding the accuracy of teachers' perceptions about low achieving learners in primary schools in rural india: An empirical analysis of alignments and misalignments. *International Journal of Educational Research Open 3*, 100198.

Wolf, S., A. Ogan, K. K. Jasińska, S. Kembou, V. K. Cannanure, H. Whitehead, J. Corbett, F. Tanoh, B. Wortsman, J. Lim, H. Zhao, and A. Kumar (2026). Impacts of targeted instruction with mobile-based support on teachers and students in rural Côte d'Ivoire. *American Educational Research Journal 63*(4), 691–730.

World Bank (2017). What is happening inside classrooms in Indian secondary schools? A time on task study in Madhya Pradesh and Tamil Nadu. Washington, DC: The World Bank and Educational Initiatives.

World Bank (2018a). *World Development Report 2018: Learning to realize education's promise*. Washington, DC: The World Bank. Washington, DC: The World Bank.

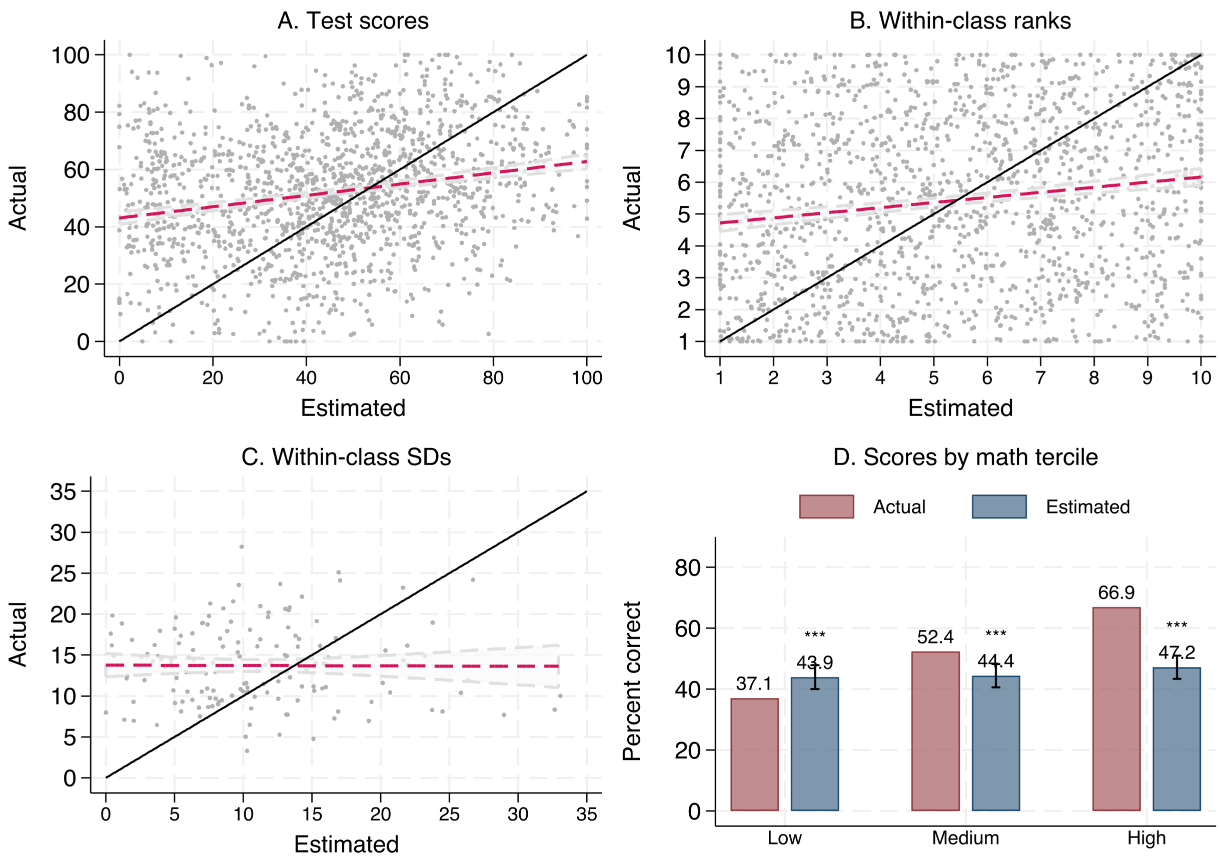
World Bank (2018b). *World Development Report 2018: Learning to realize education's promise*. Washington, DC: The World Bank.

World Bank (2023). Learning Acceleration in Secondary Education Operation (P178487). Program Appraisal Document. `https://documents1.worldbank.org/curated/en/099090523122017246/pdf/BOSIB088cf52390990af10029ec20109467.pdf`.

World Bank (2025). Learning Acceleration in Secondary Education Operation (P178487). Implementation Status & Results Report, sequence no. 5, archived December 12, 2025. https://documents1.worldbank.org/curated/en/099121225014523950/pdf/P178487-1b6e0dae-03a3-41aa-9ede-acf2f7aa620f.pdf.

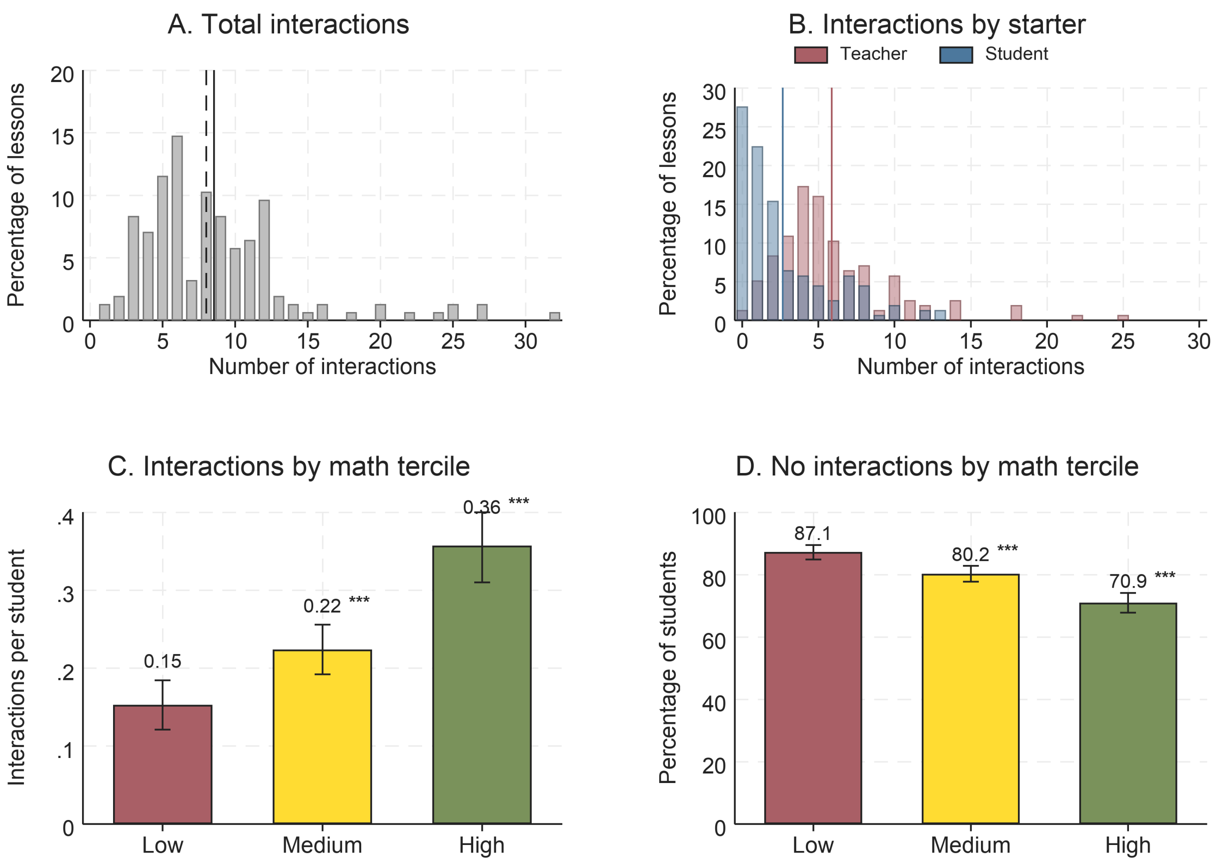
World Bank (2026). Population, total. https://data.worldbank.org/indicator/SP.POP.TOTL. Retrieved: May 21, 2026.

Figure 1: Teachers' estimations of students' achievement (baseline)



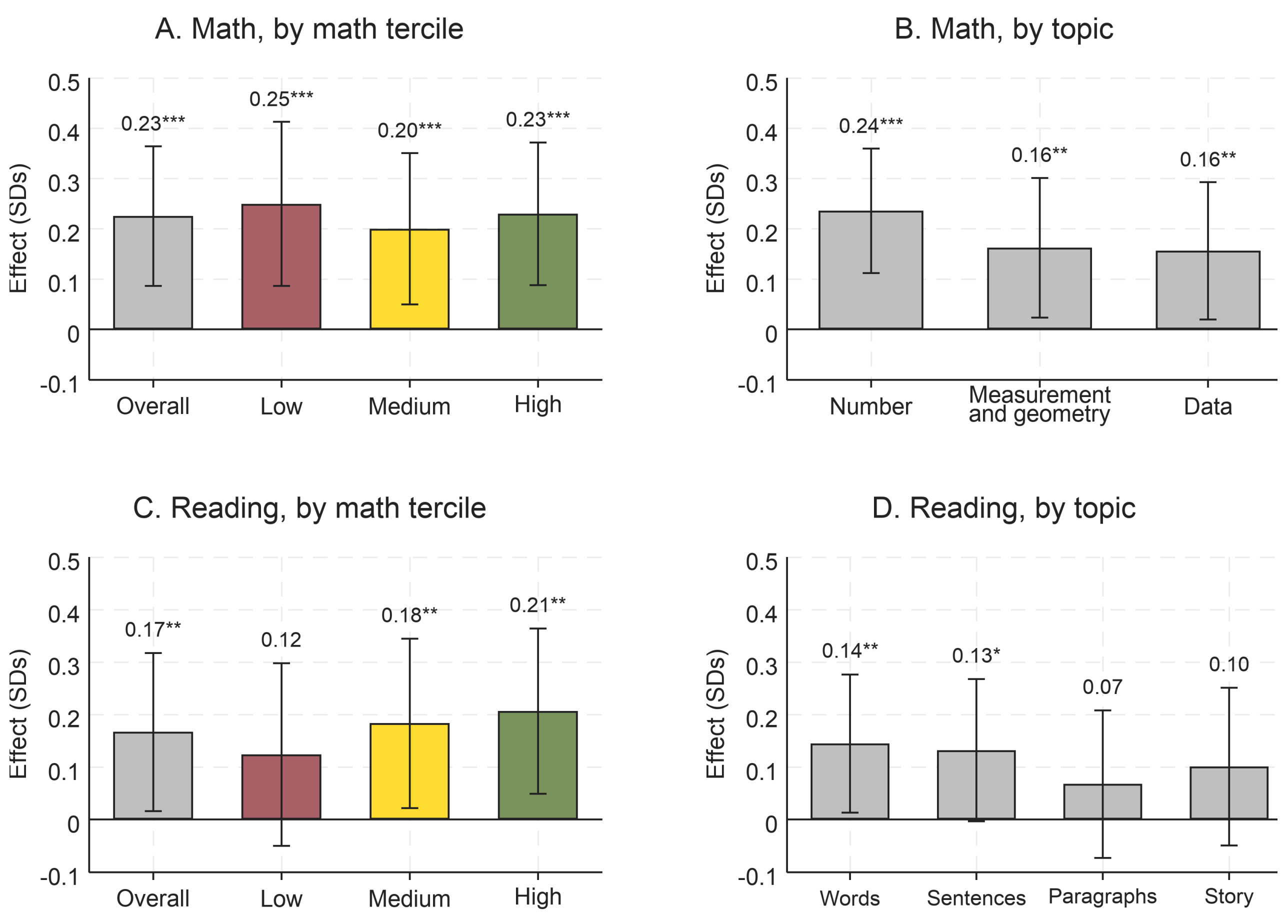
Notes: Panel A shows the relationship between students' actual intervention-assessment scores and teachers' estimated baseline math test scores. Panels B and C show the relationships between students' actual and estimated within-class ranks and between actual and estimated within-class SDs. Panel D compares actual intervention-assessment and estimated percent-correct scores by students' within-class baseline math-achievement tercile in the intervention assessments. Panels A, B, and D are at the student-teacher dyad level; panel C is at the teacher level. The red dashed line is a linear fit and the black line is the 45-degree line. In panel D, bars are strata-adjusted means, and whiskers around estimated bars show 95% confidence intervals for estimated-minus-actual differences from specifications with randomization-strata fixed effects and school-clustered standard errors. The correlations between actual and estimated values are 0.23 in panel A, 0.16 in panel B, and -0.01 in panel C. The figure includes treatment-group students with intervention-assessment scores at baseline.

Figure 2: Student-teacher interactions in the control group (endline)



Notes: Panels A and B show the distribution of student-teacher interactions per lesson in control schools. Solid lines mark means, and the dashed line in panel A marks the median. Panels C and D report adjusted means for control students by within-class baseline math-achievement tercile in the evaluation assessments from pooled control-group regressions with tercile indicators and randomization-strata fixed effects. Whiskers show 95% confidence intervals; stars indicate tests against the low-achieving group.

Figure 3: Impact on student achievement (endline)



Notes: Bars show treatment effects; whiskers show 95% confidence intervals. Panels A and C report effects overall and by within-class baseline math-achievement tercile in the evaluation assessments. Panels B and D report average effects by math topic and reading topic, respectively. All outcomes are IRT-scaled scores standardized to have a mean of zero and an SD of one in the overall control group. All specifications include randomization-strata fixed effects and baseline covariates selected via LASSO. Standard errors are clustered at the school level. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1: Change in intervention assessments by flagged content and students among treatment schools (baseline to midline)

	(1) Baseline mean	(2) Midline mean	(3) Change	(4) N (class-content/ student observations)
A. Flagged v. non-flagged topics				
Topics				
(1) Non-flagged topics	55.5	65.3	9.80	312
(2) Flagged topic	21.5	46.5	25.0	156
Row (2) − (1)			14.5***	468
			(2.71)	
Sub-topics				
(1) Non-flagged sub-topics	54.6	65.3	10.8	1,248
(2) Flagged sub-topics	4.52	42.2	37.6	156
Row (2) − (1)			19.4***	1,404
			(3.38)	
Items by color thresholds				
(1) Green items	86.0	83.0	-2.92	681
(2) Red/yellow items	25.4	51.2	25.8	879
Row (2) − (1)			34.7***	1,560
			(2.48)	
(3) Yellow items	50.6	64.8	14.2	310
(4) Red items	11.7	43.8	32.2	569
Row (4) − (3)			28.1***	879
			(2.90)	
B. Flagged v. non-flagged students				
(1) Non-flagged students	54.3	67.0	12.7	4,154
(2) Flagged students	40.7	63.3	22.5	809
Row (2) − (1)			9.85***	4,963
			(0.94)	
C. Students by number of flagged topics				
(0) 0 topics	54.3	67.0	12.7	4,154
(1) 1 topic	42.6	63.8	21.2	606
(2) 2 topics	36.4	61.9	25.5	176
(3) 3 topics	26.3	59.6	33.3	27
Row (1) − (0)			8.51***	4,963
			(1.04)	
Row (2) − (0)			12.8***	4,963
			(1.72)	
Row (3) − (0)			20.6***	4,963
			(5.08)	
D. Flagged v. non-flagged topics among flagged students				
(1) Non-flagged topics	49.4	59.2	9.79	1,388
(2) Flagged topics	13.1	54.9	41.8	958
Row (2) − (1)			32.9***	2,346
			(1.70)	

Notes: The table uses treatment classes/students with baseline and midline intervention-assessment scores. Panel A compares flagged and non-flagged topics; panel B compares flagged and non-flagged students; panel C compares students by the number of topics in which they were flagged; and panel D compares flagged and non-flagged topics among flagged students. Item colors are based on baseline means: red $<33\%$, yellow 33% to $<66\%$, and green $\geq 66\%$. Content comparisons include class and topic/sub-topic/item fixed effects; panel D uses student and topic fixed effects. Scores, changes, and differences are in pp.; school-clustered standard errors appear in parentheses. * 10%; ** 5%; *** 1%.

Table 2: Impact on math achievement (endline)

	(1) All students	(2) Low-achieving	(3) Medium-achieving	(4) High-achieving	(5) p (2)=(3)	(6) p (2)=(4)
A. Overall						
Overall	0.23***	0.25***	0.20***	0.23***	0.29	0.54
	(0.07)	(0.08)	(0.08)	(0.07)		
Control mean	0.00	-0.30	-0.01	0.32		
B. By topic						
Number	0.24***	0.27***	0.22***	0.23***	0.41	0.55
	(0.06)	(0.08)	(0.07)	(0.07)		
Control mean	0.00	-0.30	-0.01	0.32		
Measurement and geometry	0.16**	0.20**	0.12	0.17**	0.12	0.40
	(0.07)	(0.08)	(0.08)	(0.07)		
Control mean	0.00	-0.24	0.00	0.25		
Data	0.16**	0.13	0.16**	0.18**	0.67	0.55
	(0.07)	(0.08)	(0.08)	(0.08)		
Control mean	0.00	-0.17	-0.03	0.21		
p (equality across topics)	0.13					
C. By skill						
Knowing	0.18**	0.21**	0.17**	0.15**	0.43	0.30
	(0.07)	(0.08)	(0.08)	(0.07)		
Control mean	0.00	-0.27	-0.01	0.30		
Applying	0.19***	0.21***	0.17**	0.22***	0.59	0.86
	(0.06)	(0.08)	(0.07)	(0.06)		
Control mean	0.00	-0.26	-0.01	0.27		
Reasoning	0.26***	0.28***	0.24***	0.27***	0.36	0.63
	(0.07)	(0.08)	(0.08)	(0.07)		
Control mean	0.00	-0.26	-0.01	0.28		
p (equality across skills)	0.14					
N (students)	4,542	1,524	1,576	1,442		

Notes: The table shows the impact of the intervention on the math endline test, overall (panel A), by topic (panel B), and by cognitive skill (panel C). IRT-scaled scores are standardized to have a mean of zero and an SD of one in the overall control group; control means are therefore zero for column 1 by construction and negative (positive) for low- (high-) achieving students. Column 1 shows results for all students, and columns 2-4 for students in the bottom, middle, and top within-class terciles of baseline math achievement in the evaluation assessments, respectively. Columns 5 and 6 report p-values from tests of equality of the treatment effects in columns 2 and 3, and in columns 2 and 4, respectively. The p (equality across topics) and p (equality across skills) rows report omnibus tests that the three treatment effects in the corresponding panel are equal, estimated from stacked regressions with domain-specific randomization-strata fixed effects and baseline controls. All specifications use standard errors clustered at the school level; treatment-effect specifications include baseline covariates selected via LASSO. Standard errors appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: Impact on reading achievement (endline)

	(1) All students	(2) Low-achieving	(3) Medium-achieving	(4) High-achieving	(5) p (2)=(3)	(6) p (2)=(4)
A. Overall						
Reading	0.17**	0.12	0.18**	0.21**	0.46	0.39
	(0.08)	(0.09)	(0.08)	(0.08)		
Control mean	0.00	-0.19	-0.05	0.24		
B. By input						
Words						
Treatment	0.15**	0.15*	0.15*	0.13**	0.94	0.76
	(0.07)	(0.08)	(0.08)	(0.07)		
Control mean	0.00	-0.21	-0.02	0.24		
Sentences						
Treatment	0.13*	0.09	0.14*	0.18**	0.57	0.40
	(0.07)	(0.08)	(0.08)	(0.08)		
Control mean	0.00	-0.13	-0.04	0.17		
Paragraphs						
Treatment	0.07	-0.01	0.11	0.12	0.12	0.13
	(0.07)	(0.08)	(0.08)	(0.08)		
Control mean	0.00	-0.12	-0.06	0.19		
Story						
Treatment	0.10	0.07	0.09	0.16*	0.91	0.28
	(0.08)	(0.09)	(0.08)	(0.09)		
Control mean	0.00	0.00	-0.02	0.02		
C. By skill						
Vocabulary						
Treatment	0.15**	0.13	0.15*	0.16**	0.89	0.90
	(0.07)	(0.08)	(0.08)	(0.07)		
Control mean	0.00	-0.22	-0.03	0.26		
Comprehension						
Treatment	0.14*	0.07	0.17**	0.20**	0.21	0.15
	(0.08)	(0.09)	(0.08)	(0.09)		
Control mean	0.00	-0.11	-0.05	0.17		
N (students)	4,542	1,524	1,576	1,442		

Notes: The table shows the impact of the intervention on the reading endline test, overall (panel A), by input level (panel B), and by reading skill (panel C). Inputs (Words, Sentences, Paragraphs, Story) and skills (Vocabulary, Comprehension) follow the reading assessment framework, in which each item is classified by both. IRT-scaled scores are standardized to have a mean of zero and an SD of one in the overall control group; control means are therefore zero for column 1 by construction and negative (positive) for low- (high-) achieving students. Column 1 shows results for all students, and columns 2-4 for students in the bottom, middle, and top within-class terciles of baseline math achievement in the evaluation assessments, respectively. Columns 5 and 6 report p-values from tests of equality of the treatment effects in columns 2 and 3, and in columns 2 and 4, respectively. All specifications include randomization-strata fixed effects and baseline covariates selected via LASSO. Standard errors clustered at the school level appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: Impact on social-emotional skills (endline)

	(1) All students	(2) Low-achieving	(3) Medium-achieving	(4) High-achieving	(5) p (2)=(3)	(6) p (2)=(4)
Self-efficacy	0.14***	0.14**	0.18***	0.09	0.51	0.53
	(0.05)	(0.07)	(0.06)	(0.07)		
Self-management	0.05	-0.01	0.09	0.06	0.15	0.36
	(0.05)	(0.07)	(0.06)	(0.07)		
Growth mindset	0.11**	0.20***	0.09	0.04	0.16	0.03
	(0.05)	(0.07)	(0.06)	(0.06)		
Fear of failure	-0.03	0.02	-0.07	-0.04	0.19	0.44
	(0.05)	(0.07)	(0.06)	(0.07)		
Task-goal orientation	0.10**	0.13*	0.10	0.07	0.65	0.42
	(0.05)	(0.07)	(0.06)	(0.06)		
Motivation to master tasks	0.03	0.03	0.06	-0.02	0.69	0.46
	(0.05)	(0.06)	(0.06)	(0.06)		
Grit	0.06	0.07	0.09	0.01	0.83	0.35
	(0.05)	(0.07)	(0.06)	(0.07)		
N (students)	4,542	1,524	1,576	1,442		

Notes: The table shows the impact of the intervention on seven self-reported social-emotional skills measured at endline. Graded-response IRT models pool the baseline and endline items; scores are standardized to have a mean of zero and an SD of one in the control group in each round. Column 1 shows results for all students, and columns 2-4 for students in the bottom, middle, and top within-class terciles of baseline math achievement in the evaluation assessments, respectively. Columns 5 and 6 report p-values from tests of equality of the treatment effects in columns 2 and 3, and in columns 2 and 4, respectively. All specifications include randomization-strata fixed effects and baseline covariates selected via LASSO. Standard errors clustered at the school level appear in parentheses. * 10%; ** 5%; *** 1%.

Table 5: Impact on teachers' estimations of students' test scores (endline)

	(1) All students	(2) Low-achieving	(3) Medium-achieving	(4) High-achieving	(5) p (2)=(3)	(6) p (2)=(4)
A. Accuracy						
Absolute score gap (SDs)	0.14	0.11	0.20	0.09	0.50	0.90
	(0.20)	(0.19)	(0.24)	(0.19)		
Control mean	1.74	1.83	1.77	1.62		
Absolute rank gap (percentiles)	-0.01	0.01	-0.01	-0.02	0.42	0.08
	(0.01)	(0.01)	(0.01)	(0.01)		
Control mean	0.31	0.31	0.31	0.29		
B. Direction						
Score gap (SDs)	-0.47*	-0.48*	-0.48*	-0.42*	1.00	0.69
	(0.24)	(0.25)	(0.29)	(0.23)		
Control mean	-0.26	-0.11	-0.33	-0.35		
Rank gap (percentiles)	-0.01*	-0.01	-0.01	-0.01	0.96	0.95
	(0.01)	(0.02)	(0.02)	(0.02)		
Control mean	0.00	0.04	0.00	-0.04		
Teacher overestimates	-0.04	-0.03	-0.04	-0.05	0.88	0.61
	(0.04)	(0.04)	(0.04)	(0.04)		
Control mean	0.46	0.50	0.46	0.40		
Teacher underestimates	0.03	0.03	0.04	0.03	0.72	1.00
	(0.04)	(0.04)	(0.04)	(0.04)		
Control mean	0.50	0.46	0.50	0.56		
C. Dispersion						
SD gap (points)	0.07	0.00	-0.01	0.17	0.97	0.50
	(0.72)	(0.73)	(0.74)	(0.73)		
Control mean	-0.63	-0.50	-0.71	-0.69		
N (student-teacher dyads)	4,314	1,423	1,503	1,388		

Notes: Impact on teachers' estimations of students' math endline performance. Panel A shows absolute score gaps (normalized by the within-class SD of actual scores) and absolute rank gaps (within-class percentile rank). Panel B shows signed score and rank gaps and indicators for over- and under-estimation. Panel C shows the SD gap. Negative (positive) signed gaps indicate under- (over-) estimation. Cols. 2-4 are within-class terciles of baseline math achievement in the evaluation assessments; cols. 5-6 give p-values for cols. 2 vs. 3 and 2 vs. 4. All specs include strata FEs and LASSO-selected baseline covariates. School-clustered SEs in parentheses. * 10%; ** 5%; *** 1%.

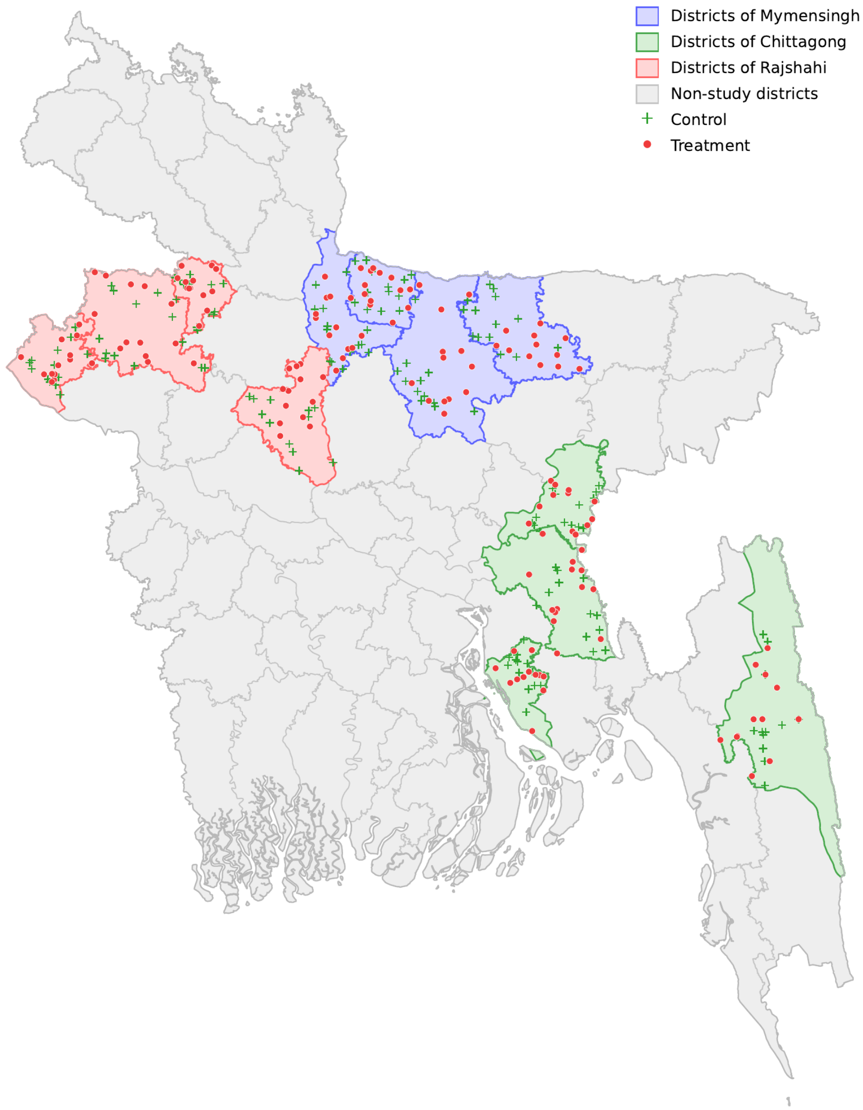
Table 6: Impact on student-teacher interactions (endline)

	(1) All students	(2) Low-achieving	(3) Medium-achieving	(4) High-achieving	(5) p (2)=(3)	(6) p (2)=(4)
A. Overall						
Number of interactions	-0.02	0.01	-0.02	-0.07*	0.22	0.04
	(0.02)	(0.02)	(0.02)	(0.04)		
Control mean	0.24	0.15	0.22	0.36		
Share with 0 interactions	0.02	-0.02	0.02	0.05**	0.10	0.02
	(0.01)	(0.02)	(0.02)	(0.02)		
Control mean	0.80	0.87	0.80	0.71		
B. By starter						
Teacher	-0.02	0.00	-0.01	-0.04	0.67	0.39
	(0.02)	(0.02)	(0.02)	(0.03)		
Control mean	0.16	0.11	0.15	0.24		
Student	0.00	0.02	-0.01	-0.03	0.15	0.02
	(0.01)	(0.01)	(0.01)	(0.02)		
Control mean	0.08	0.04	0.07	0.12		
C. By type						
Statement	-0.01	0.00	0.00	-0.04	0.95	0.14
	(0.01)	(0.02)	(0.01)	(0.02)		
Control mean	0.08	0.06	0.06	0.13		
Question	-0.01	0.02	-0.01	-0.03	0.11	0.13
	(0.02)	(0.02)	(0.02)	(0.03)		
Control mean	0.16	0.10	0.16	0.23		
D. By length						
$\leq$ 5 seconds	-0.01	0.01	0.00	-0.03*	0.27	$<$ 0.01
	(0.01)	(0.01)	(0.01)	(0.02)		
Control mean	0.04	0.02	0.04	0.06		
6-10 seconds	0.00	0.01	0.00	-0.01	0.50	0.57
	(0.01)	(0.01)	(0.01)	(0.02)		
Control mean	0.08	0.05	0.08	0.12		
$\geq$ 10 seconds	-0.02	-0.01	-0.01	-0.03	0.62	0.41
	(0.01)	(0.01)	(0.02)	(0.03)		
Control mean	0.12	0.08	0.11	0.17		
N (students)	5,920	2,077	2,058	1,785		

Notes: The table shows the impact of the intervention on the number of student-teacher interactions per student per lesson observed at endline, overall (panel A), by who started the interaction (panel B), by interaction type (panel C), and by interaction length (panel D). Outcomes are observed counts per student during a single classroom-observation lesson. Column 1 shows results for all students, and columns 2-4 for students in the bottom, middle, and top within-class terciles of baseline math achievement in the evaluation assessments, respectively. Columns 5 and 6 report p-values from tests of equality of the treatment effects in columns 2 and 3, and in columns 2 and 4, respectively. All specifications include randomization-strata fixed effects and baseline covariates selected via LASSO. Standard errors clustered at the school level appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

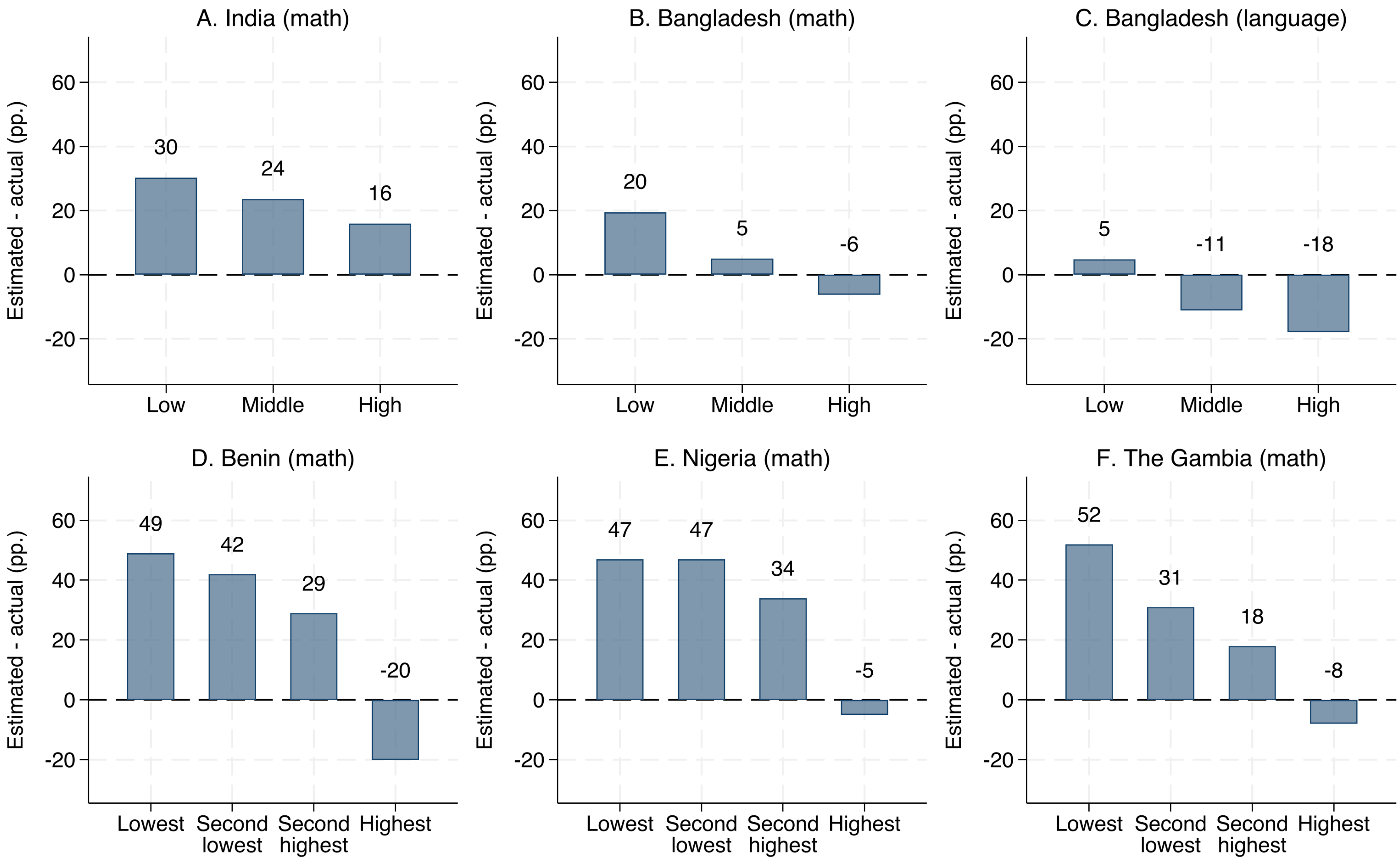
Appendix A Additional figures and tables

Figure A.1: Spatial distribution of study schools



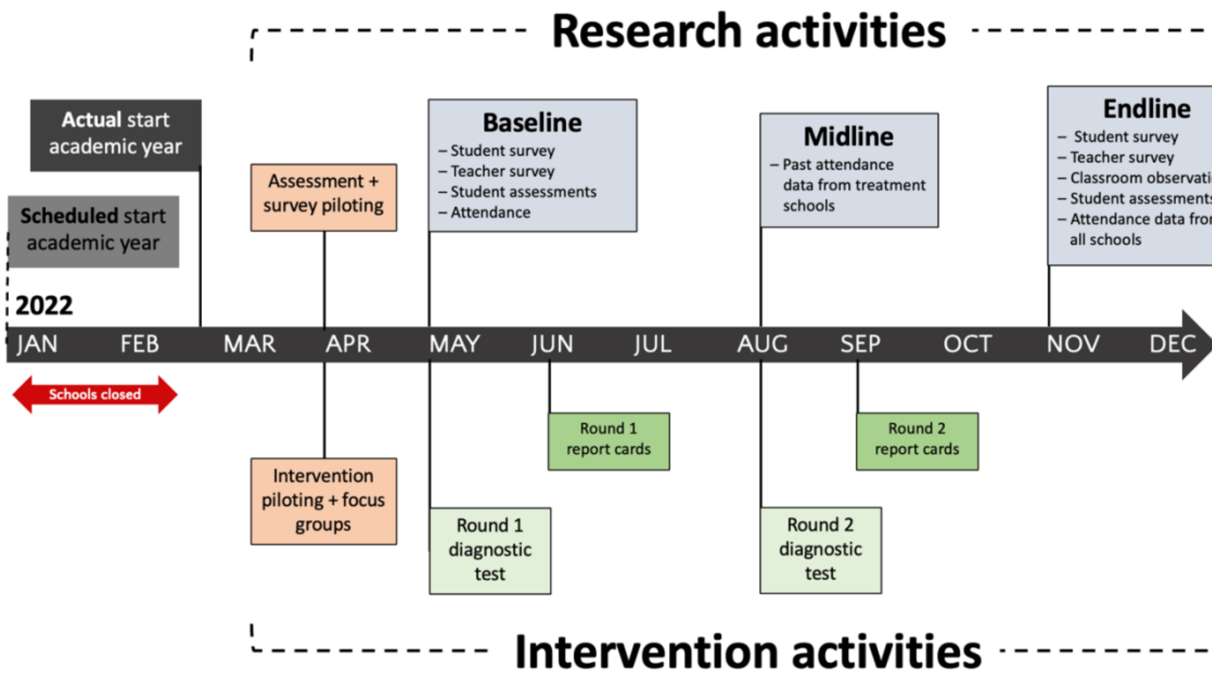
Notes: The figure shows the geo-tagged location of all schools in the impact evaluation across Chattogram (formerly Chittagong) in green, Mymensingh in blue, and Rajshahi in red.

Figure A.2: External evidence on teachers' estimations of students' achievement (baseline)



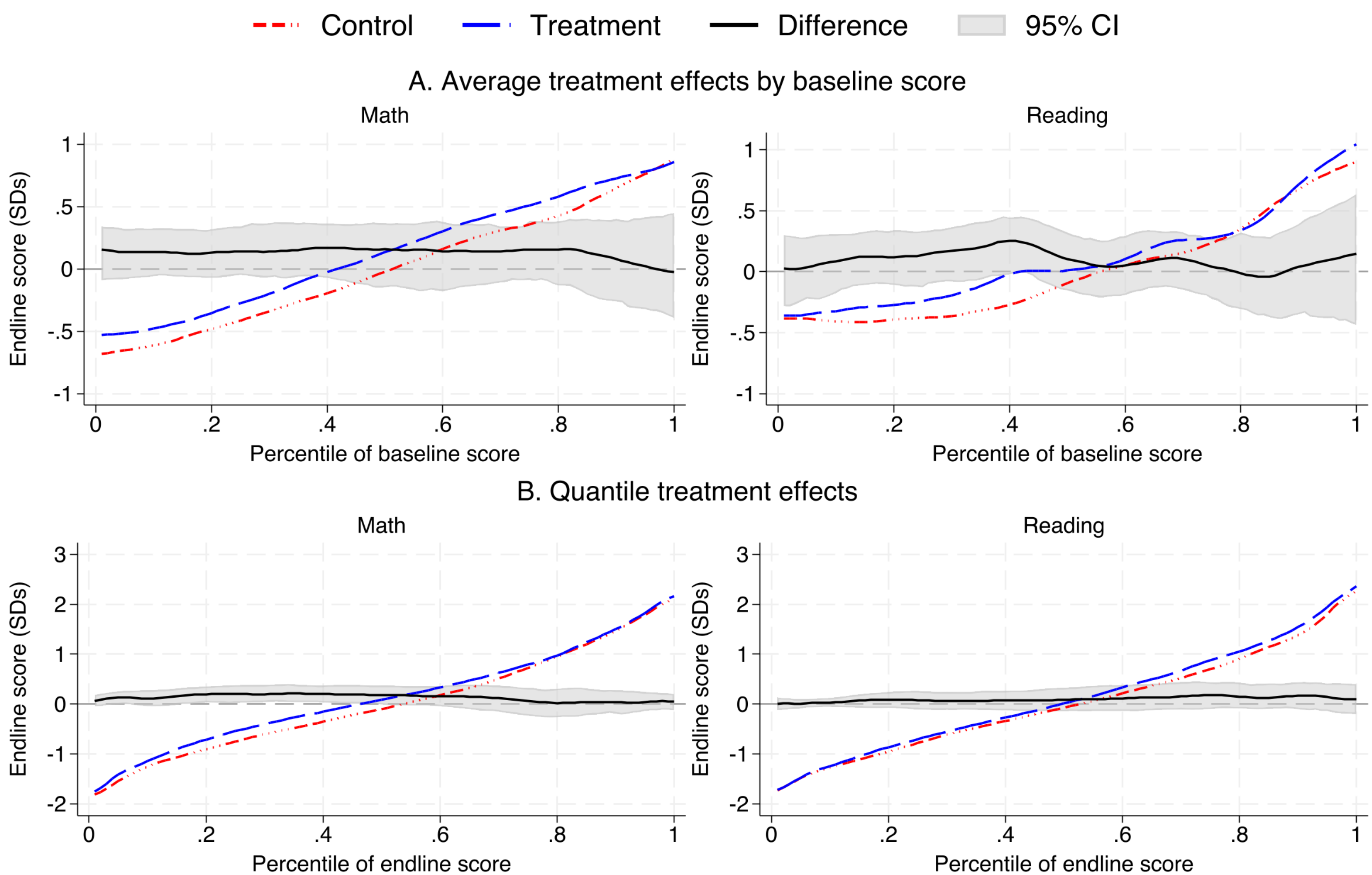
Notes: Panels A-C show estimated-minus-actual gaps in students' percentage-correct scores by students' within-class achievement tercile in India math, Bangladesh math, and Bangladesh language, using values from Table A.6 in Djaker et al. (2024). Panels D-F show estimated-minus-actual gaps in headteachers' beliefs about the percentage of students who could answer an addition problem correctly, by school-performance quartile, in Benin, Nigeria, and The Gambia, using graphs from Rodriguez-Segura (2024). Positive values indicate overestimation.

Figure A.3: Timeline of research and intervention activities



Notes: The figure shows the timeline for research and intervention activities during the 2022 school year. Schools reopened on February 22, 2022, after closures due to COVID-19. March and April were used to pilot the intervention, assessments, and surveys.

Figure A.4: Distributional treatment effects on student achievement (endline)



Notes: The figure shows distributional treatment effects on endline math (left column) and reading (right column) achievement. In each panel, the red and blue lines are local-polynomial regressions of standardized endline scores on the within-group percentile rank of either baseline scores (top row) or endline scores (bottom row), estimated separately for control and treatment students; the black line is their difference, and the shaded band is its 95% confidence interval, constructed via a 200-iteration cluster bootstrap (school clusters, stratified by treatment). Scores are standardized to have a mean of zero and an SD of one in the control group.

Table A.1: Comparison between in- and out-of-sample schools on national student assessment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All divisions				In-sample divisions			
	All schools	Out-of-sample schools	In-sample schools	Col.(3)-(2)	All schools	Out-of-sample schools	In-sample schools	Col.(7)-(6)
A. Math								
Total (std.)	0.00	0.00	-0.19	-0.20	0.04	0.05	-0.19	-0.25
	[0.95]	[0.95]	[0.81]	(0.15)	[0.96]	[0.96]	[0.81]	(0.15)
Numbers (pct.)	0.61	0.61	0.58	-0.04	0.62	0.63	0.58	-0.05
	[0.23]	[0.23]	[0.20]	(0.03)	[0.23]	[0.23]	[0.20]	(0.03)
Measurement (pct.)	0.55	0.56	0.50	-0.06	0.57	0.57	0.50	-0.07
	[0.27]	[0.27]	[0.26]	(0.04)	[0.28]	[0.28]	[0.26]	(0.04)
Data (pct.)	0.64	0.64	0.61	-0.02	0.64	0.64	0.61	-0.03
	[0.26]	[0.26]	[0.27]	(0.04)	[0.26]	[0.26]	[0.27]	(0.04)
Geometry (pct.)	0.52	0.53	0.45	-0.07*	0.53	0.54	0.45	-0.08**
	[0.25]	[0.25]	[0.23]	(0.04)	[0.25]	[0.25]	[0.23]	(0.04)
Algebra (pct.)	0.48	0.48	0.48	-0.01	0.49	0.49	0.48	-0.02
	[0.28]	[0.28]	[0.25]	(0.04)	[0.28]	[0.28]	[0.25]	(0.04)
N (students)	28,566	27,802	470	28,566	10,926	10,456	470	10,926
B. Bangla								
Total (std.)	0.00	0.00	-0.14	-0.14	0.03	0.04	-0.14	-0.18
	[0.94]	[0.94]	[0.90]	(0.12)	[0.91]	[0.91]	[0.90]	(0.13)
Reading (pct.)	0.63	0.63	0.60	-0.03	0.64	0.64	0.60	-0.03
	[0.19]	[0.19]	[0.19]	(0.02)	[0.18]	[0.18]	[0.19]	(0.03)
Vocabulary (pct.)	0.81	0.81	0.80	-0.01	0.81	0.81	0.80	-0.02
	[0.20]	[0.20]	[0.20]	(0.02)	[0.20]	[0.20]	[0.20]	(0.02)
Grammar (pct.)	0.78	0.78	0.75	-0.04	0.79	0.79	0.75	-0.04*
	[0.22]	[0.22]	[0.22]	(0.02)	[0.22]	[0.22]	[0.22]	(0.03)
N (students)	28,566	27,802	470	28,566	10,926	10,456	470	10,926

Notes: The table shows the means and SDs of all out-of-sample schools and in-sample schools, and the difference between these groups on Bangladesh's national student assessment, the National Assessment of Secondary Students (NASS). Columns 1-4 report national comparisons, while columns 5-8 restrict the sample to the three divisions included in the experiment. Panel A shows results for math and panel B for Bangla (language). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.2: Balancing checks on student characteristics and skills (baseline)

Variable	(1) Control	(2) Treatment	(3) Col.(2)-(1)
A. Background			
Female	0.61	0.65	0.04*
	[0.49]	[0.48]	(0.03)
Age	12.0	12.0	-0.04
	[0.99]	[0.95]	(0.04)
Speaks Bangla at home	0.94	0.94	0.00
	[0.24]	[0.23]	(0.01)
Mother completed high school	0.45	0.40	-0.05***
	[0.50]	[0.49]	(0.02)
Household assets (std.)	0.00	0.13	0.14***
	[1.00]	[0.99]	(0.05)
N (students)	2,966	2,974	5,940
B. Social-emotional skills			
Growth mindset (std.)	0.00	0.06	0.05
	[1.00]	[1.00]	(0.04)
Self management (std.)	0.00	0.02	0.02
	[1.00]	[0.99]	(0.04)
Self efficacy (std.)	0.00	0.07	0.07
	[1.00]	[0.98]	(0.04)
Task-goal orientation (std.)	0.00	0.00	0.00
	[1.00]	[0.95]	(0.03)
Motivation to master tasks (std.)	0.00	-0.01	-0.01
	[1.00]	[0.99]	(0.04)
Fear of failure (std.)	0.00	-0.03	-0.03
	[1.00]	[1.03]	(0.04)
Grit (std.)	0.00	0.03	0.03
	[1.00]	[1.02]	(0.04)
N (students)	2,966	2,974	5,940
C. Academic skills			
Math (IRT, std.)	0.00	-0.17	-0.18**
	[1.00]	[1.01]	(0.07)
Reading (IRT, std.)	0.00	-0.13	-0.14*
	[1.00]	[1.00]	(0.08)
Fluid intelligence (std.)	0.00	-0.06	-0.06
	[1.00]	[1.03]	(0.05)
N (students)	2,966	2,974	5,940

Notes: The table shows the means and SDs of students in the control (column 1) and treatment groups (column 2). It also tests for differences between groups, accounting for randomization-strata fixed effects (column 3). Panel A shows results for background characteristics, panel B for social-emotional skills, and panel C for academic skills. The p-value from a joint F-test of all listed covariates in a regression of treatment status on those covariates, with randomization-strata fixed effects and standard errors clustered by school, is < 0.001. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Balancing checks on teacher characteristics and skills (baseline)

Variable	(1) Control	(2) Treatment	(3) Col.(2)-(1)
A. Background			
Female	0.12	0.15	0.03
	[0.32]	[0.36]	(0.04)
Age	41.7	41.8	0.04
	[8.67]	[7.64]	(0.93)
N (teachers)	156	155	311
B. Education and experience			
Education major	0.07	0.09	0.02
	[0.26]	[0.29]	(0.03)
Math major	0.31	0.28	-0.03
	[0.47]	[0.45]	(0.05)
Has master's degree	0.31	0.36	0.05
	[0.47]	[0.48]	(0.05)
Has pre-service training	0.28	0.18	-0.10**
	[0.45]	[0.39]	(0.05)
Has in-service training	0.47	0.52	0.06
	[0.50]	[0.50]	(0.06)
Experience teaching	14.5	15.3	0.85
	[9.16]	[9.13]	(1.03)
Experience at this school	12.9	13.9	1.00
	[9.27]	[8.75]	(1.03)
N (teachers)	156	155	311
C. Social-emotional skills			
Self efficacy (std.)	0.00	0.11	0.11
	[1.00]	[0.90]	(0.11)
Job satisfaction (std.)	0.00	0.01	0.01
	[1.00]	[0.93]	(0.11)
Locus of control (std.)	0.00	0.10	0.10
	[1.00]	[1.04]	(0.11)
Grit (std.)	0.00	0.01	0.00
	[1.00]	[1.01]	(0.11)
Growth mindset (std.)	0.00	0.06	0.06
	[1.00]	[1.14]	(0.12)
N (teachers)	156	155	311

Notes: The table shows the means and SDs of teachers in the control (column 1) and treatment groups (column 2). It also tests for differences between groups, accounting for randomization-strata fixed effects (column 3). Panel A shows results for background characteristics, panel B for education and experience, and panel C for social-emotional skills. The p-value from a joint F-test of all listed covariates in a regression of treatment status on those covariates, with randomization-strata fixed effects and standard errors clustered by school, is 0.499. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Attrition in student assessments (endline)

	(1) Attrited
A. Treatment	
Treatment	0.02
	(0.02)
N (students)	5940
Control mean	0.23
B. Treatment and baseline	
Treatment	0.01
	(0.15)
Female	-0.06***
	(0.02)
Age	0.05***
	(0.01)
Composite test score (std.)	-0.03***
	(0.01)
Treatment × Female	-0.01
	(0.03)
Treatment × Age	0.00
	(0.01)
Treatment × Composite score	0.01
	(0.01)
N (students)	5921
F-ratio (Interactions)	0.75
P-value	0.53

Notes: The table shows estimates from regressions predicting follow-up status in the endline assessments. Follow-up is defined as having an observed test score at endline. The sample includes students present at baseline. Panel A regresses attrition on treatment status, and panel B regresses attrition on treatment status interacted with baseline characteristics. Both panels include randomization-strata fixed effects. Standard errors (clustered by school) appear in parentheses. The F- and p-values refer to a test of joint significance for all interaction terms. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: Teachers' estimations of students' achievement on concrete tasks (baseline)

	(1) Actual	(2) Estimated	(3) Difference
A. Percentage of proficient students, by topic			
Number	5.40	41.8	36.5***
			(1.40)
Algebra	18.1	52.1	34.0***
			(1.60)
Measurement	0.30	36.0	35.6***
			(1.30)
Geometry	2.70	40.8	38.1***
			(1.50)
Data	0.90	45.7	44.8***
			(1.50)
N (teachers)	311	311	311
B. Percentage of students answering specific items correctly			
Operations	56.9	39.6	-17.3***
			(2.10)
Geometry	30.6	42.7	12.2***
			(1.90)
Data	48.3	50.8	2.60
			(2.00)
N (teachers)	311	311	311
C. Probability that specific students answer specific items correctly			
Operations	0.82	0.90	0.08***
			(0.02)
Geometry	0.24	0.54	0.30***
			(0.02)
N (student-teacher-item responses)	3,063	3,063	3,063

Notes: The table shows actual and estimated student performance on concrete math tasks. Panel A shows the actual and estimated percentage of "proficient" students (i.e., those who can answer most questions correctly) on each topic at the teacher level. Panel B shows the actual and estimated percentage of students who can answer three items correctly at the teacher level: in operations, "What is the missing sign in 72_6 + 15 = 27?"; in measurement and geometry, "Name a triangle with two equal sides"; and in data, "Find the mode of 8, 12, 11, 4, 7, 12, 62." Panel C shows the actual and estimated probability that specific students can answer two items correctly at the student-teacher-item level: in operations, "What is 15×4?"; and in measurement and geometry, "A rectangular box is 6 cm long, 4 cm wide, and 2 cm thick. What is the volume of the box?" Differences equal estimated minus actual and are estimated from regressions of actual and estimated values on an indicator for estimated values, with randomization-strata fixed effects. Standard errors clustered at the school level appear in parentheses. The table includes both experimental groups at baseline. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6: Stallings classroom observations (endline)

	(1) Control	(2) Treatment	(3) Col.(2)-(1)
A. Lesson time allocation (%)			
On task	93.6	93.3	-0.24
	[10.3]	[10.9]	(1.17)
Class management	1.76	1.84	0.08
	[4.59]	[5.46]	(0.54)
Off task	4.65	4.81	0.16
	[9.55]	[9.89]	(1.07)
B. Lesson time on task, by activity (%)			
Reading aloud	9.38	9.30	-0.08
	[19.9]	[18.0]	(1.92)
Explanation	22.2	22.2	0.00
	[21.4]	[19.6]	(2.09)
Demonstration	10.5	11.7	1.20
	[19.4]	[17.5]	(1.82)
Discussion	22.8	23.6	0.80
	[22.9]	[24.3]	(2.48)
Practice and drill	11.1	9.38	-1.68
	[17.2]	[15.2]	(1.76)
Assignment	6.89	7.13	0.24
	[11.5]	[12.1]	(1.30)
Copying	3.13	3.37	0.24
	[7.99]	[8.91]	(0.92)
Instructions	7.61	6.65	-0.96
	[12.4]	[11.7]	(1.28)
C. Student engagement (% of lesson time)			
No students	5.45	6.09	0.64
	[11.0]	[11.3]	(1.20)
1 student	6.81	7.53	0.72
	[12.6]	[13.3]	(1.41)
2-10 students	7.29	8.89	1.60
	[12.8]	[14.6]	(1.51)
10+ students	11.8	11.5	-0.24
	[21.5]	[21.1]	(2.21)
All students	68.7	65.9	-2.72
	[26.9]	[28.5]	(3.00)
N (schools)	156	156	312

Notes: The table presents results from the endline Stallings-style classroom observations. Panel A shows the percentage of lesson time allocated to on-task activities, class management, and off-task activities. Panel B decomposes on-task activities as a percentage of lesson time. Panel C shows student engagement as a percentage of lesson time. Each school contributes one observation. Columns 1 and 2 report the means and SDs (in brackets) of each outcome in the control and treatment groups, respectively; column 3 reports the difference, estimated by a regression on the treatment indicator with randomization-strata fixed effects. Heteroskedasticity-robust standard errors appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.7: Student-teacher interactions by baseline achievement in the control group (endline)

	(1) All students	(2) Low- achieving	(3) Medium- achieving	(4) High- achieving	(5) p (2)=(3)	(6) p (2)=(4)
A. Overall						
Number of interactions	0.24	0.15	0.22	0.36	< 0.01	< 0.01
	(0.01)	(0.02)	(0.02)	(0.02)		
Share with 0 interactions	0.80	0.87	0.80	0.71	< 0.01	< 0.01
	(0.01)	(0.01)	(0.01)	(0.02)		
B. By starter						
Teacher	0.16	0.11	0.15	0.24	0.01	< 0.01
	(0.01)	(0.01)	(0.01)	(0.02)		
Student	0.08	0.04	0.07	0.12	0.01	< 0.01
	(0.01)	(0.01)	(0.01)	(0.01)		
C. By type						
Statement	0.08	0.06	0.06	0.13	0.57	< 0.01
	(0.01)	(0.01)	(0.01)	(0.01)		
Question	0.16	0.10	0.16	0.23	< 0.01	< 0.01
	(0.01)	(0.01)	(0.02)	(0.02)		
D. By direction						
Unidirectional	0.03	0.02	0.02	0.05	0.59	0.01
	(0.01)	(0.01)	(0.01)	(0.01)		
Bidirectional	0.16	0.11	0.16	0.23	< 0.01	< 0.01
	(0.01)	(0.01)	(0.01)	(0.02)		
Multidirectional	0.05	0.03	0.04	0.08	0.20	< 0.01
	(0.01)	(0.01)	(0.01)	(0.01)		
E. By content						
Academic	0.22	0.14	0.21	0.34	< 0.01	< 0.01
	(0.01)	(0.02)	(0.02)	(0.02)		
Logistic	0.01	0.01	0.01	0.01	0.11	0.69
	(0.00)	(0.00)	(0.00)	(0.00)		
Social	0.01	0.00	0.00	0.01	0.43	0.12
	(0.00)	(0.00)	(0.00)	(0.00)		
F. By length						
< 5 seconds	0.04	0.02	0.04	0.06	0.06	< 0.01
	(0.01)	(0.01)	(0.01)	(0.01)		
6-10 seconds	0.08	0.05	0.08	0.12	0.01	< 0.01
	(0.01)	(0.01)	(0.01)	(0.01)		
> 10 seconds	0.12	0.08	0.11	0.17	0.12	< 0.01
	(0.01)	(0.01)	(0.01)	(0.02)		
N (students)	2,964	1,041	1,030	893		

Notes: The table shows adjusted means for the number of student-teacher interactions per student per lesson observed at endline in control schools, overall (panel A), by who started the interaction (panel B), by interaction type (panel C), by interaction direction (panel D), by content (panel E), and by interaction length (panel F). Outcomes are observed counts per student during a single classroom-observation lesson. Column 1 shows results for all control students, and columns 2-4 for control students in the bottom, middle, and top within-class terciles of baseline math achievement in the evaluation assessments, respectively. Columns 5 and 6 report p-values from tests that the adjusted means in columns 2 and 3, and in columns 2 and 4, are equal. All specifications restrict the sample to control-group students and include randomization-strata fixed effects. Standard errors clustered at the school level appear in parentheses.

Table A.8: Student-teacher interactions by teacher characteristics in the control group (endline)

	(1)	(2)	(3)	(4)	(5)	(6)
	Yes/ $\geq$ median		No/ $<$ median		Col. (1)-(3)	
	Mean	SD	Mean	SD	Diff.	SE
A. Demographics						
Female	7.56	[4.37]	8.67	[5.40]	-0.69	(0.64)
Age	8.73	[5.15]	8.33	[5.48]	1.02*	(0.53)
B. Education						
Education major	7.18	[3.66]	8.65	[5.39]	-3.41**	(1.68)
Math major	8.61	[4.21]	8.51	[5.74]	0.83	(0.60)
Has master's degree	7.20	[3.35]	9.16	[5.89]	-0.59	(0.55)
Has pre-service training	7.86	[4.19]	8.81	[5.66]	0.30	(0.63)
Has in-service training	8.63	[5.98]	8.47	[4.64]	0.25	(0.61)
C. Experience						
Experience teaching	8.93	[5.25]	8.20	[5.33]	0.97*	(0.54)
Experience at this school	8.89	[5.31]	8.24	[5.29]	1.00*	(0.54)
D. Teacher estimations						
Gap in estimations	8.61	[5.85]	8.43	[4.85]	-0.03	(0.59)
E. Social-emotional skills						
Index (std.)	8.32	[5.22]	8.80	[5.40]	-0.42	(0.59)
Self-efficacy	7.38	[3.96]	9.78	[6.19]	-1.17**	(0.53)
Job satisfaction	8.52	[5.47]	8.60	[4.92]	-0.22	(0.62)
Locus of control	8.70	[5.28]	7.90	[5.38]	1.06	(0.65)
Grit	8.76	[5.27]	8.16	[5.35]	0.67	(0.61)
Growth mindset	9.28	[6.03]	7.78	[4.29]	0.18	(0.52)
N (teachers)	156					

Notes: The table shows the number of student-teacher interactions per lesson observed at endline in control schools, by teachers' baseline characteristics. For binary characteristics, columns 1 and 2 show teachers with the characteristic and columns 3 and 4 show teachers without it. For continuous characteristics, columns 1 and 2 show teachers at or above the median and columns 3 and 4 show teachers below the median. The index is the standardized first principal component of self-efficacy, job satisfaction, locus of control, grit, and growth mindset. Columns 1-4 report means and SDs in brackets. Columns 5 and 6 report differences estimated from regressions with randomization-strata fixed effects. Standard errors clustered at the school level appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.9: Student-teacher interactions by student characteristics in the control group (endline)

	(1)	(2)	(3)	(4)	(5)	(6)
	Yes/ ≥ median		No/ < median		Col. (1)-(3)	
	Mean	SD	Mean	SD	Diff.	SE
A. Background						
Female	0.23	[0.53]	0.25	[0.59]	0.02	(0.02)
Age	0.22	[0.54]	0.29	[0.59]	-0.06***	(0.02)
Speaks Bangla at home	0.21	[0.49]	0.68	[1.09]	0.20	(0.24)
Mother completed high school	0.26	[0.52]	0.22	[0.58]	0.09***	(0.02)
Household assets (std.)	0.28	[0.62]	0.19	[0.46]	0.10***	(0.02)
B. Social-emotional skills						
Index (std.)	0.28	[0.59]	0.19	[0.52]	0.09***	(0.02)
Growth mindset (std.)	0.28	[0.58]	0.18	[0.52]	0.10***	(0.02)
Self management (std.)	0.26	[0.58]	0.21	[0.52]	0.05**	(0.02)
Self efficacy (std.)	0.28	[0.59]	0.19	[0.50]	0.09***	(0.02)
Task-goal orientation (std.)	0.24	[0.53]	0.24	[0.62]	0.01	(0.02)
Motivation to master tasks (std.)	0.26	[0.55]	0.21	[0.57]	0.06***	(0.02)
Fear of failure (std.)	0.24	[0.56]	0.23	[0.56]	0.00	(0.03)
Grit (std.)	0.22	[0.51]	0.26	[0.61]	-0.04**	(0.02)
C. Academic skills						
Index (std.)	0.31	[0.64]	0.16	[0.45]	0.15***	(0.02)
Math (IRT, std.)	0.31	[0.63]	0.16	[0.46]	0.15***	(0.02)
Reading (IRT, std.)	0.29	[0.62]	0.19	[0.48]	0.10***	(0.02)
Fluid intelligence (std.)	0.27	[0.59]	0.19	[0.50]	0.08***	(0.02)
N (students)	2,964					

Notes: The table shows the number of student-teacher interactions per student during the endline classroom observation in control schools, by students' baseline characteristics. For binary characteristics, columns 1 and 2 show students with the characteristic and columns 3 and 4 show students without it. For continuous characteristics, columns 1 and 2 show students at or above the class-level median and columns 3 and 4 show students below the class-level median. The social-emotional index is the standardized first principal component of growth mindset, self-management, self-efficacy, fear of failure (reverse-coded), goal orientation, motivation to master tasks, and grit. The academic index is the standardized first principal component of math achievement, reading achievement, and fluid intelligence. Columns 1-4 report means and SDs in brackets. Columns 5 and 6 report differences estimated from regressions with randomization-strata fixed effects. Standard errors clustered at the school level appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.10: Student-teacher interactions and teachers' estimation accuracy in the control group (endline)

	(1) All students	(2) Low-achieving	(3) Medium-achieving	(4) High-achieving	(5) p (2)=(3)	(6) p (2)=(4)
Score gap (SDs)	0.17*	0.03	0.29**	0.21**	0.08	0.15
	(0.09)	(0.10)	(0.14)	(0.10)		
Absolute score gap (SDs)	-0.18**	-0.11	-0.20	-0.14*	0.46	0.75
	(0.09)	(0.09)	(0.14)	(0.08)		
Rank gap (percentiles)	0.03*	0.01	0.08***	0.03	0.10	0.55
	(0.02)	(0.03)	(0.03)	(0.02)		
Absolute rank gap (percentiles)	-0.02*	-0.01	0.00	-0.03**	0.58	0.46
	(0.01)	(0.02)	(0.02)	(0.01)		
Teacher overestimates	0.01	-0.01	0.02	0.04	0.55	0.15
	(0.02)	(0.02)	(0.03)	(0.03)		
N (student-teacher dyads)	2,189	722	759	708		

Notes: The table shows coefficients from regressions of teacher-estimation outcomes on the number of observed interactions between the teacher and student during the endline classroom observation. Column 1 uses all control-group students; columns 2-4 use control-group students in the bottom, middle, and top within-class terciles of baseline math achievement in the evaluation assessments, respectively. Columns 5 and 6 report p-values from tests that the coefficients in columns 2 and 3, and in columns 2 and 4, are equal. Score gaps equal estimated minus actual scores, scaled by the within-class SD of actual scores; rank gaps equal estimated minus actual within-class percentile ranks; absolute gaps are the absolute values of those measures. All specifications restrict the sample to control-group students and include school fixed effects. Standard errors clustered at the school level appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.11: Impact on math achievement by flagged status (endline)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Flagged		Non-flagged		Difference		Equality p-value	
	Report-based	Evaluation-based	Report-based	Evaluation-based	Report-based	Evaluation-based	Report-based	Evaluation-based
Overall math	0.21***	0.24***	0.23***	0.21***	-0.02	0.03	0.76	0.57
	(0.08)	(0.08)	(0.08)	(0.07)	(0.06)	(0.05)		
Number	0.19*	0.26***	0.25***	0.23***	-0.05	0.03	0.54	0.60
	(0.10)	(0.08)	(0.06)	(0.06)	(0.09)	(0.06)		
Measurement and geometry	0.14	0.21**	0.17**	0.15**	-0.02	0.06	0.78	0.32
	(0.09)	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)		
Data	0.12	0.17**	0.16**	0.15**	-0.04	0.02	0.65	0.80
	(0.10)	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)		
N (students)				4,543				

Notes: The table shows impacts on endline math achievement by flagged status. Within each column pair, report-based columns use observed report flags for treatment students and the bottom three students within each class on each baseline evaluation-assessment topic for control students; evaluation-based columns use this bottom-three rule in both groups. Columns 1-2 report effects for flagged students; columns 3-4 report effects for non-flagged students; columns 5-6 report the differences between these effects; and columns 7-8 report p-values from tests of their equality. IRT-scaled scores are standardized to have a mean of zero and an SD of one in the overall control group. All specifications include randomization-strata fixed effects and baseline covariates selected via LASSO. Standard errors clustered at the school level appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.12: Impact on reading achievement by math-flagged status (endline)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Flagged		Non-flagged		Difference		Equality p-value	
	Report-based	Evaluation-based	Report-based	Evaluation-based	Report-based	Evaluation-based	Report-based	Evaluation-based
Overall reading	0.15*	0.13	0.20**	0.20**	-0.05	-0.07	0.45	0.13
	(0.09)	(0.08)	(0.08)	(0.08)	(0.07)	(0.05)		
Words	0.12	0.10	0.18**	0.18***	-0.06	-0.08	0.35	0.12
	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.05)		
Sentences	0.13	0.12	0.14*	0.15**	-0.01	-0.03	0.87	0.58
	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)		
Paragraphs	0.04	0.03	0.10	0.10	-0.06	-0.08	0.41	0.19
	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.06)		
Story	0.16*	0.07	0.12	0.13	0.04	-0.06	0.58	0.35
	(0.09)	(0.07)	(0.08)	(0.09)	(0.07)	(0.06)		
Vocabulary	0.12	0.11	0.18**	0.18**	-0.06	-0.08	0.33	0.14
	(0.09)	(0.08)	(0.07)	(0.07)	(0.07)	(0.05)		
Comprehension	0.15*	0.10	0.17**	0.17**	-0.02	-0.07	0.83	0.21
	(0.09)	(0.08)	(0.08)	(0.08)	(0.07)	(0.06)		
N (students)	4,543							

Notes: The table shows impacts on endline reading achievement by math-flagged status. Within each column pair, report-based columns use observed report flags for treatment students and the bottom three students within each class on each baseline evaluation-assessment math topic for control students; evaluation-based columns use this bottom-three rule in both groups. Columns 1-2 report effects for flagged students; columns 3-4 report effects for non-flagged students; columns 5-6 report the differences between these effects; and columns 7-8 report p-values from tests of their equality. IRT-scaled scores are standardized to have a mean of zero and an SD of one in the overall control group. All specifications include randomization-strata fixed effects and baseline covariates selected via LASSO. Standard errors clustered at the school level appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.13: Impact on teachers' topic-specific estimations of students' test scores (endline)

	(1) All students	(2) Low-achieving	(3) Medium-achieving	(4) High-achieving	(5) p (2)=(3)	(6) p (2)=(4)
A. Gap in % of proficient students, by topic (pp.)						
Number	4.30	3.92	4.24	4.27	0.76	0.70
	(2.92)	(2.96)	(2.97)	(2.96)		
Control mean	41.6	42.1	41.5	41.3		
Measurement and geometry	0.94	0.43	1.75	0.99	0.10	0.48
	(2.51)	(2.55)	(2.57)	(2.52)		
Control mean	46.5	47.1	46.3	46.0		
Data	3.69	3.88*	4.13*	3.07	0.72	0.26
	(2.24)	(2.27)	(2.32)	(2.25)		
Control mean	65.2	65.6	64.5	65.5		
B. Gap in % of students answering specific items correctly, by topic (pp.)						
Operations	-1.01	-1.27	-0.52	-1.40	0.53	0.91
	(3.24)	(3.27)	(3.40)	(3.25)		
Control mean	-5.50	-5.23	-5.56	-5.69		
Geometry	1.23	0.27	2.00	1.39	0.14	0.30
	(3.11)	(3.22)	(3.14)	(3.15)		
Control mean	-1.45	-0.29	-2.14	-1.91		
Data	-2.81	-3.86	-2.05	-2.79	0.10	0.37
	(3.42)	(3.49)	(3.48)	(3.36)		
Control mean	15.6	16.5	14.8	15.5		
N (student-teacher dyads)	4,314	1,423	1,503	1,388		

Notes: Appendix table for Table 5. The table shows the impact of the intervention on the gap between estimated and actual % proficient (panel A) and % answering specific items correctly (panel B), by topic. Negative (positive) coefficients indicate under- (over-) estimation. Cols. 2-4 are within-class terciles of baseline math achievement in the evaluation assessments; cols. 5-6 give p-values for cols. 2 vs. 3 and 2 vs. 4. All specs include strata FEs and LASSO-selected baseline covariates. School-clustered SEs in parentheses. * 10%; ** 5%; *** 1%.

Table A.14: Impact on topics taught during observed lessons (endline)

	(1) Control	(2) Treatment	(3) Col.(2)-(1)
Number	0.73	0.86	0.14
	[0.45]	[0.35]	(0.09)
Measurement and geometry	0.14	0.14	-0.01
	[0.35]	[0.35]	(0.08)
Data	0.18	0.04	-0.15**
	[0.39]	[0.20]	(0.07)
N (schools)	44	49	93

Notes: The table shows the impact of the intervention on whether the observed lesson covered each topic, measured at endline in a random sub-sample of schools. Each school contributes one observation. Columns 1 and 2 report the means and SDs (in brackets) of each outcome in the control and treatment groups, respectively. Column 3 reports the difference, estimated by a regression on the treatment indicator with randomization-strata fixed effects; heteroskedasticity-robust standard errors are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.15: Impact on student and teacher attendance (midline and endline)

	(1) All students	(2) Low-achieving	(3) Medium-achieving	(4) High-achieving	(5) p (2)=(3)	(6) p (2)=(4)
A. Student attendance						
School registers						
Midline	-0.01	-0.01	-0.02	-0.01	0.10	0.74
	(0.02)	(0.02)	(0.02)	(0.02)		
Control mean	0.70	0.68	0.71	0.73		
N (students)	5,935	2,084	2,063	1,787		
Endline	0.02	0.04**	-0.01	0.02	< 0.01	0.22
	(0.01)	(0.02)	(0.02)	(0.02)		
Control mean	0.71	0.66	0.72	0.74		
N (students)	5,917	2,077	2,059	1,780		
Observed attendance						
Midline	0.01					
	(0.02)					
Control mean	0.62					
N (schools)	311					
Endline	0.02					
	(0.02)					
Control mean	0.61					
N (schools)	310					
B. Teacher attendance						
School registers						
Midline	0.01*					
	(0.01)					
Control mean	0.93					
N (teachers)	304					
Endline	0.00					
	(0.01)					
Control mean	0.93					
N (teachers)	306					

Notes: The table shows impacts on student attendance (panel A) and teacher attendance (panel B) at midline and endline. School-register attendance is calculated as days present divided by working days. Observed attendance is calculated as the number of students present on the survey day divided by enrollment. Column 1 reports effects for all students and pooled teacher effects. Columns 2-4 report effects on register-based student attendance for students in the bottom, middle, and top within-class terciles of baseline math achievement in the evaluation assessments, respectively. Columns 5 and 6 report p-values from tests of equality of the treatment effects in columns 2 and 3, and in columns 2 and 4, respectively. Control means and sample sizes appear below each set of estimates. All specifications include randomization-strata fixed effects; observed-attendance specifications also control for baseline observed attendance. Standard errors clustered at the school level appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix B Intervention components

The intervention comprised three components, described in turn below.

B.1 Intervention assessments

Students in treatment schools completed two short, 10-item math assessments during the 2022 school year. The first was administered in May and the second in August (Figure A.3).[32] We decided on the distribution of items across topics (number, measurement and geometry, and data) and skills (knowing, applying, and reasoning) based on the assessment framework for the Trends in International Mathematics and Science Study (TIMSS), an international student assessment of math and science for grades 4 and 8 administered by the International Association for the Evaluation of Educational Achievement (IEA, 2021). We then adjusted the framework based on the Bangladeshi national curriculum (NCTB, 2023). The assessment framework for the intervention assessments is shown in Tables B.1 and B.2.

B.2 Reports for teachers

After each assessment, teachers received a brief, five-page learning report (Figures B.1-B.5). Each report summarized the purpose of the exercise, the class's overall performance, students' performance on each item, performance by topic, the students most in need of support in each topic, the operations that appeared to need reinforcement, and concrete suggestions for using the results in instruction. Reports used traffic-light colors for item-level results: red when no more than 32% of students answered correctly, yellow when 33 to 65% did, and green when at least 66% did. Reports emphasized narrative descriptions of results rather than graphical displays, and their format was refined through three rounds of feedback with Bangladeshi teachers to make the information clearer, easier to interpret, and more useful.

B.3 Error-analysis guide

Teachers also received an error-analysis guide that linked common incorrect answers to potential misconceptions. For each item, the guide identified the correct answer, described the process needed to obtain it, and interpreted the misconceptions most likely behind each incorrect response (Figures B.6 and B.7).

[32]The school year in Bangladesh had begun in February after COVID-related closures; March and April were used to pilot the tests and data-collection instruments.

Figure B.1: Learning-report template, page 1: report overview and purpose

Government Secondary School

EIIN # 99999

REPORT CARD ON STUDENT LEARNING

Why are you receiving this report card?

The Government of Bangladesh and the World Bank have partnered to provide schools with data on student learning to help you identify students in need of additional support.

What information does it include?

The report card summarizes the performance of your students in foundational math skills, based on assessments conducted in August 2022 at your school. It also compares their performance to that of students attending similar schools in your district and shows improvements of your students from June 2022.

Why is this important?

Students need to master foundational math skills to understand the grade 6 curriculum. Although you may not have been responsible for teaching students these skills, it is important that you address any gaps in students' knowledge to help them better understand your lessons.

How should you use it?

- Identify the *areas* (e.g., numbers and operations) and *topics* (e.g., subtraction) in which your students are performing least well;
- Prepare one or more lessons covering the areas and topics that need reinforcement, checking whether students understand the material during class;
- Reach out to the students who perform least well in these areas and topics according to this report, review their homework, and ask them if they need additional help;
- Track the progress of the low-performing students using the last round's report card that you received in June 2022

How will you know if your students have improved?

We visted your school in June 2022 and gave you a similar report, comparing with this will give you an idea of your students' progress.

Notes: The figure shows the first page of the five-page learning report provided to treatment teachers after each intervention assessment. The page introduces the report, states its purpose, summarizes how it was constructed, and explains how teachers should interpret the information that follows.

Figure B.2: Learning-report template, page 2: performance summary and item scores

How did your students perform compared to others in your district?

- The typical student in your class answered **6 out of 10** (60%) of the questions correctly, compared to **6 out of 10** (60%) questions in your district.
- In your class, **31%** of your students are **low-performers**, **35%** are **medium performers**, and **11%** are **high-performers**, compared to **37%, 35%, and 27%** in your district.
- Low-performing students in your class answered **4 out of 10** (40%) questions correctly, compared to **4 out of 10** (40%) in your district. These are the students that need your support the most.
- The area in which students in your class performed least well was **measurements and geometry**.
- The topics in which students in your class need most help were **addition of numbers and conversion of measurement units**, and **applying geometric properties to find missing angle**.

How did your students perform in each question?

Content area	Question-number in the test, % of correct answers in this question				
Numbers and operations	Q1 60%	Q2 85%	Q3 88%	Q4 84%	Q5 95%
Geometry and measurements	Q7 3%	Q8 39%	Q9 12%		
Data	Q6 65%	Q10 39%			

Q1: Addition of a whole number and a decimal number.
Q2: Multiplication of a one-digit number by a two-digit number.
Q3: Word problem involving a subtraction of a one-digit number by another.
Q4: Division of two-digit number by a one-digit number.
Q5: Word problem involving cross multiplication.
Q6: Interpreting a calendar to make extrapolations.
Q7: Addition of numbers and conversion of measurement units.
Q8: Calculating the volume of a three-dimensional shape.
Q9: Applying geometric properties to find missing angle.
Q10: Interpreting tables with data to calculate a fraction.

Notes: The figure shows the second page of the learning report. It reports the class's overall performance on the assessment and the share of students who answered each item correctly, allowing teachers to identify items that the class found particularly difficult.

Figure B.3: Learning-report template, page 3: performance by topic and students in need of support

Performance in each topic:

Numbers and operations

- The typical student in your class answered **4 out of 5** questions correctly in this area.
- Within this area, your students struggle the most with **addition of a whole number and a decimal number** and **division of two-digit number by a one-digit number**.
- The three students in your class who need the most help in this area are:
  1. **Miton Das (Roll No. 18, Male)**
  2. **Titas Das (Roll No. 23, Male)**
  3. **Md Mehedi Hasan (Roll No. 41, Male)**

Geometry and measurements

- The typical student in your class answered **1 out of 3** questions correctly in this area.
- Within this area, your students struggle the most with **addition of numbers and conversion of measurement units**.
- The three students in your class who need the most help in this area are:
  1. **Mst Takiya Akter (Roll No. 28, Female)**
  2. **Israt Jahan Thrisha (Roll No. 64, Female)**
  3. **Md Abjal (Roll No. 29, Male)**

Data

- The typical student in your class answered **1 out of 2** questions correctly in this area.
- Within this area, your students struggle the most with **interpreting tables with data to calculate a fraction**.
- The three students in your class who need the most help in this area are:
  1. **Mst Takiya Akter (Roll No. 28, Female)**
  2. **Mst Maisha Akter Maya (Roll No. 7, Female)**
  3. **Mst Tachlima Khandaker (Roll No. 66, Female)**

Notes: The figure shows the third page of the learning report. It breaks down class performance by topic (number, measurement and geometry, and data) and names the students most in need of support on each topic.

Figure B.4: Learning-report template, page 4: improvement summary and operations needing reinforcement

Improvements from June 2022:

How did your students improve?

- The typical student in your class answered 1 more questions out of 10 correctly.
- In your class, **60%** answered fewer questions correctly, **16%** answered more questions correctly, and **24%** answered the same number of questions correctly.
- Low-performing students in your class answered 1 more questions out of 10 correctly.
- The area in which students in your class improved the least was **measurements and geometry**.
- The topics in which students in your class improved the least were **applying geometric properties to find missing angle**, and **addition of numbers and conversion of measurement units**.

Improvement in each topic:

Numbers and operations

- The typical student in your class answered 1 more questions out of 5 correctly in this area.
- Within this area, your students struggle the most with **division of two-digit number by a one-digit number** and **division of two-digit number by a one-digit number**.

Geometry and measurements

- The typical student in your class answered the same number of questions out of 3 correctly in this area.
- Within this area, your students struggle the most with **applying geometric properties to find missing angle**.

Data

- The typical student in your class answered the same number of questions out of 2 correctly in this area.
- Within this area, your students struggle the most with **interpreting a calendar to make extrapolations**.

Notes: The figure shows the fourth page of the learning report. It summarizes how class performance changed since the previous assessment (where applicable) and flags specific mathematical operations and procedures that appeared to need reinforcement.

Figure B.5: Learning-report template, page 5: potential uses of the report

What should you do next?

Here is what you should do next:

1. Use the lists on the previous page to identify the students in need of most help;
2. Use the additional copies of the test administered in May 2022 and ask these students to attempt it once again;
3. Review each student's answers and identify the questions he/she answered incorrectly;
4. Ask each student to explain to you what they thought the question was asking, how they approached finding a solution, and whether they can think of another way to solve it.
5. Use the error-analysis document to explain to students why the option that they selected is incorrect and why another option is correct instead.

Here are some tips on how to do it:

- Students may be embarrassed of their mistakes, so reassure them that you are asking them to redo the test to **understand what they got wrong and how they can improve**.
- The objective of this process is to understand what the student does not know, so **resist the temptation to help them** when they are re-taking the test.
- If a student gets stuck, **ask him/her what are the options** that he/she is considering. If he/she remains silent, **offer him/her a hint**, but ask him/her whether he/she knows why that hint may be correct.

Notes: The figure shows the fifth and final page of the learning report. It offers concrete suggestions for how teachers can use the information in the report—e.g., to group students by need, to revisit specific operations in upcoming lessons, or to communicate with families about progress.

Figure B.6: Error-analysis guide, example 1: addition of decimal numbers

1. What is sum of 11.2 and 25?
   a. 3.62
   b. 13.6
   c. 13.8
   d. 36.2

The correct answer is D: 36.2

- ***First, the student should add the numbers in the ones place value (i.e., 1 from 11 and 5 from 25) and write down the result (i.e., 6).***
- ***Then, he/she should add the numbers in the tens place value (i.e., 1 from 11 and 2 from 25) and write down the result (i.e., 3).***
- ***Then, he/she should add the numbers in the decimals (i.e., 2 from 11.2 and 0 from 25.0) and write down the result (i.e., 2).***
- ***With a 3 in the tens place value, a 6 in the ones place value, and 2 in decimal, the answer is 36.2***

If a student chose...

a. He/she could have could have performed the calculation correctly but made an error in understanding how decimals should be considered in addition operations.

b. He/she could have misrecognized operation and added 11 and 2 from the first number, while also making an error in decimals.

c. He/she could have performed the calculation correctly except with misaligning the numbers without considering the digits in the decimals.

Notes: The figure shows the first example from the error-analysis guide provided to treatment teachers alongside the assessment results. For each item, the guide lists the correct answer, the procedure required to reach it, and the misconceptions most likely to underlie each incorrect response. Similar explanations were provided for all 10 items in each assessment.

Figure B.7: Error-analysis guide, example 2: word problem involving subtraction

3. Marie has 9 pencils. She gives 4 pencils to her sister. How many pencils does Marie have now?
    a. 9
    b. 13
    c. 4
    d. 5

The correct answer is D: 5.

- ***First, the student should recognize that this is a subtraction question with 4 being subtracted from 9. He/she should note the order of subtraction and should not be confused in it.***
- ***Then, the student should subtract 4 from 9 and record the answer 5.***
- ***The student should write down the correct answer as 5.***

If a student chose…

a. He/she could have simply copied down the number 9 directly from the question without recognizing any operation.
b. He/she could have misinterpreted the operation to be addition instead of subtraction.
c. He/she could have simply copied down the number 4 directly from the question without recognizing any operation.

Notes: The figure shows the second example from the error-analysis guide provided to treatment teachers alongside the assessment results, illustrating how the guide addressed contextual word problems in addition to procedural items.

Table B.1: Math intervention assessment framework, by topic

(1) Topic	(2) Target	(3) Actual	(4) Topics assessed	(5) Items
A. Number (target: 50%; actual: 50%)				
Whole numbers	25%	30%	• Multiply and divide whole numbers, including computation in simple contextual problems.	2, 4
			• Solve problems involving odd and even numbers, multiples and factors of numbers, rounding numbers, and making estimates.	3
Expressions, simple equations, and relationships	15%	10%	• Identify or write expressions or number sentences to represent problem situations that may involve unknowns.	5
Fractions and decimals	10%	10%	• Demonstrate knowledge of decimal place value, including representing decimals using words, numbers, or models; compare, order, and round decimals; add and subtract decimals.	1
B. Measurement and geometry (target: 30%; actual: 30%)				
Measurement	15%	10%	• Solve problems involving mass, volume, and time; identify appropriate types and sizes of units and read scales.	7
Geometry	15%	20%	• Identify and draw parallel and perpendicular lines; identify and draw right angles and angles smaller or larger than a right angle; compare angles by size.	8
			• Use elementary properties, including line and rotational symmetry, to describe, compare, and create common two-dimensional shapes.	9
C. Data (target: 20%; actual: 20%)				
Reading, interpreting, and representing data	15%	10%	• Read and interpret data from tables, pictographs, bar graphs, line graphs, and pie charts.	10
Using data to solve problems	5%	10%	• Use data to answer questions that go beyond directly reading data displays, including solving problems, combining data from multiple sources, and drawing conclusions.	6

Notes: The table shows the target and actual distribution of items by topic in the 10-item math intervention assessments administered in May and August. The framework drew on Bangladesh's grades 1-5 curriculum and an international assessment framework for primary-school mathematics. Items are referenced by their position in the assessment.

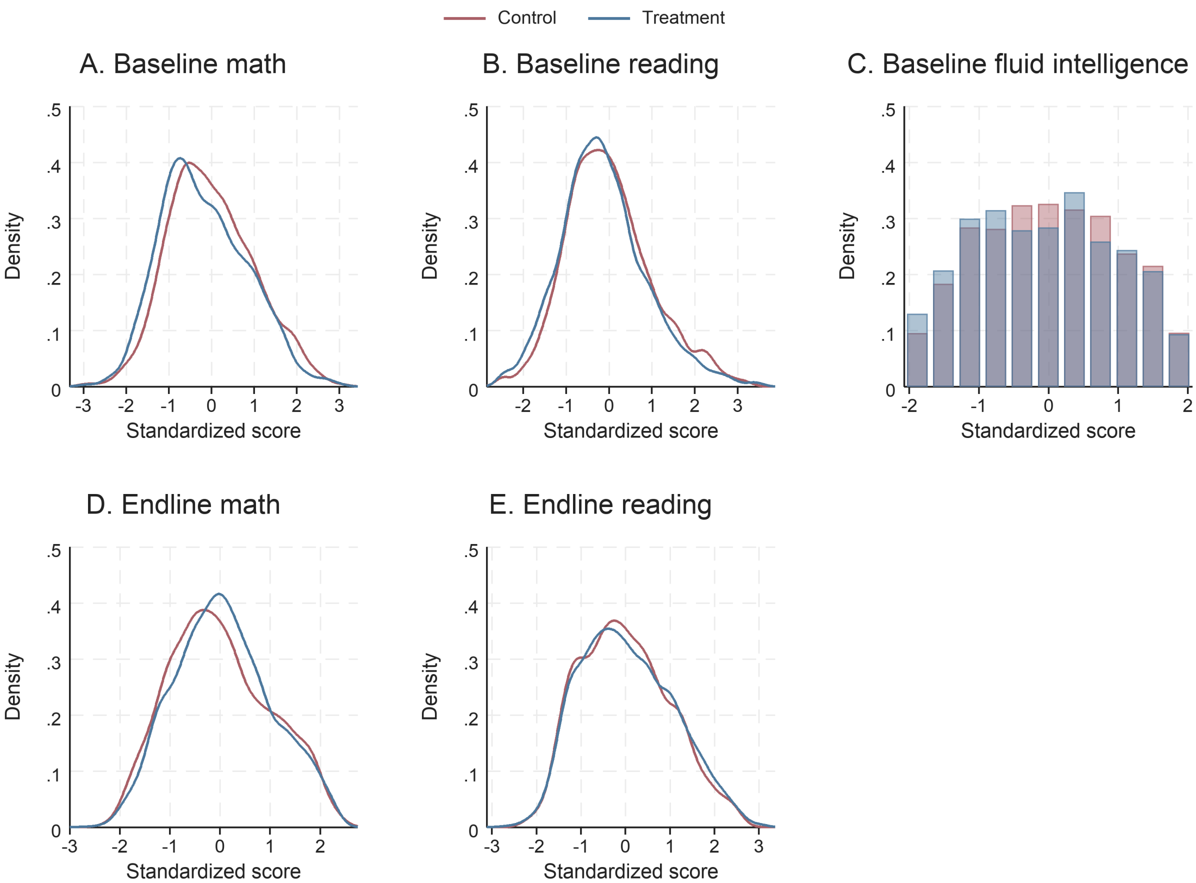
Table B.2: Math intervention assessment framework, by cognitive domain

(1) Cognitive domain	(2) Target	(3) Actual	(4) Skills assessed	(5) Items
A. Knowing (target: 40%; actual: 40%)				
Recall	10%	10%	• Recall definitions, terminology, number properties, units of measurement, geometric properties, and notation.	9
Compute	30%	30%	• Carry out algorithmic procedures for addition, subtraction, multiplication, division, or a combination of these with whole numbers, fractions, decimals, and integers; carry out straightforward algebraic procedures.	1, 2, 4
B. Applying (target: 40%; actual: 40%)				
Determine	10%	10%	• Determine efficient or appropriate operations, strategies, and tools for solving problems for which there are commonly used methods of solution.	8
Implement	30%	30%	• Implement strategies and operations to solve problems involving familiar mathematical concepts and procedures.	3, 7, 10
C. Reasoning (target: 20%; actual: 20%)				
Analyze	10%	10%	• Determine, describe, or use relationships among numbers, expressions, quantities, and shapes.	5
Integrate/Synthesize	10%	10%	• Link different elements of knowledge, related representations, and procedures to solve problems.	6

Notes: The table shows the target and actual distribution of items by cognitive skill in the 10-item math intervention assessments. Sub-skill targets are computed from the share of items in each sub-skill (the framework specifies targets only at the skill level—Knowing, Applying, and Reasoning—so we set sub-skill targets equal to their realized item share). Items are referenced by their position in the assessment.

Appendix C Measurement

Figure C.1: Student assessment distributions (baseline and endline)



Notes: The figure shows kernel density plots of math and reading scores and histograms of fluid-intelligence scores, separately by experimental group. Panels A–C show baseline scores; panels D and E show endline scores. Math and reading are IRT-scaled and standardized to the control-group distribution within round; fluid-intelligence scores are standardized to the baseline control-group distribution.

Table C.1: Math evaluation assessment framework, by topic

(1) Topic	(2) Target	(3) Actual	(4) Number of items	(5) Items
A. Number (target: 50%; actual: 47%)				
Whole numbers	25%	27%	8	1-8
Place value	0%	0%	0	–
Add and subtract	10%	10%	3	1, 3, 4
Multiply and divide	7%	7%	2	2, 6
Odd/even numbers, factors, rounding, and estimates	3%	3%	1	5
Number and operation properties	7%	7%	2	7, 8
Expressions, simple equations, and relationships	15%	13%	4	9-12
Missing number or operation	3%	3%	1	9
Expressions or number sentences	3%	3%	1	10
Patterns and relationships	7%	7%	2	11, 12
Fractions and decimals	10%	7%	2	13, 14
Fractions	3%	3%	1	13
Decimals	3%	3%	1	14
B. Measurement and geometry (target: 30%; actual: 30%)				
Measurement	15%	17%	5	15-19
Lengths	7%	7%	2	15, 19
Mass, volume, time, units, and scales	7%	7%	2	16, 17
Perimeter, area, and volume	3%	3%	1	18
Geometry	15%	13%	4	20-23
Lines and angles	3%	3%	1	20
Shape properties and symmetry	3%	3%	1	21
Patterns and relationships	7%	7%	2	22, 23
C. Data (target: 20%; actual: 23%)				
Reading, interpreting, and representing data	15%	17%	5	24-28
Read and interpret data displays	7%	7%	2	24, 25
Organize and represent data	10%	10%	3	26-28
Using data to solve problems	5%	7%	2	29, 30
Use data beyond direct reading	7%	7%	2	29, 30

Notes: The table shows the target and actual distribution of items in the baseline and endline math evaluation assessments, by topic. Target shares are from the assessment framework when available; otherwise, subdomain target shares are set equal to realized item shares because the framework does not specify separate targets for those subdomains. Actual shares and item positions are computed from the item lists in the baseline and endline framework workbooks, which use the same allocation. Items are referenced by their position in each assessment.

Table C.2: Math evaluation assessment framework, by cognitive domain

(1) Cognitive domain	(2) Target	(3) Actual	(4) Number of items	(5) Items
A. Knowing (target: 40%; actual: 43%)				
Recall	13%	13%	4	3, 12, 21, 24
Recognize	7%	7%	2	6, 20
Classify/order	0%	0%	0	–
Compute	13%	13%	4	1, 2, 14, 29
Retrieve	3%	3%	1	26
Measure	7%	7%	2	17, 18
B. Applying (target: 40%; actual: 37%)				
Determine	20%	20%	6	5, 7, 9, 13, 27, 30
Represent/model	7%	7%	2	22, 23
Implement	10%	10%	3	4, 15, 16
C. Reasoning (target: 20%; actual: 20%)				
Analyze	10%	10%	3	10, 25, 28
Integrate/synthesize	3%	3%	1	11
Evaluate	7%	7%	2	8, 19
Draw conclusions	0%	0%	0	–
Generalize	0%	0%	0	–
Justify	0%	0%	0	–

Notes: The table shows the target and actual distribution of items in the baseline and endline math evaluation assessments, by cognitive domain. Top-level target shares are from the assessment framework; sub-skill target shares are set equal to realized item shares because the framework does not specify separate targets for individual sub-skills. Actual shares and item positions are computed from the item lists in the baseline and endline framework workbooks, which use the same allocation. Items are referenced by their position in each assessment.

Table C.3: Reading assessment framework, by reading skill

(1) Domain/area	(2) Target	(3) Actual	(4) Skills assessed	(5) Items
A. Vocabulary (target: 50%; actual: 50%)				
Picture matching		20%	Identify pictures matching given words or short sentences; measures receptive-language skills.	1-6
Cloze task		3%	Sentence with one word partially or completely covered; students use the context of the sentence to identify the missing word.	7
Maze task		27%	Sentence with two or three alternative words; students select the word that makes the most sense in the sentence.	12-15, 19-22
B. Comprehension (target: 50%; actual: 50%)				
Focus on and retrieve		10%	Focus on and retrieve explicitly stated information from the text.	16-18
Make inferences		13%	Make straightforward inferences from information in the text.	8-11
Interpret and integrate		10%	Interpret and integrate ideas and information across the text.	28-30
Evaluate and critique		17%	Evaluate and critique content and textual elements.	23-27

Notes: The table shows the target and actual distribution of items by reading skill in the 30-item baseline and endline reading assessments. Top-level targets (Vocabulary and Comprehension) are specified at the domain level; the framework does not specify separate targets for individual sub-skills, so the Target column for sub-skills is left blank. Items are referenced by their position in the assessment.

Table C.4: Reading assessment framework, by input level

(1) Input level	(2) Target	(3) Actual	(4) Description	(5) Items
Words	20%	20%	Items present a single word or short phrase to which students must respond (e.g., matching a word to a picture).	1-6
Sentences	20%	20%	Items present one or a few sentences to which students must respond (e.g., completing a sentence, matching a sentence to a picture).	7, 16-18, 23-24
Paragraphs	40%	40%	Items follow a short paragraph (typically several sentences) on which students answer comprehension or cloze questions.	8-15, 19-22
Story	20%	20%	Items follow a longer narrative or informational text spanning multiple paragraphs.	25-30

Notes: The table shows the target and actual distribution of items by input level in the 30-item reading assessments. Input level refers to the length and structure of the text on which the item is based, ranging from a single word to a multi-paragraph story. Items are referenced by their position in the assessment.

Table C.5: Variance components of student-teacher interaction scores, full sample (endline)

	(1) Initiator	(2) Direction	(3) Content	(4) Tone	(5) Framing	(6) Length
Variance component	Var.	Var.	Var.	Var.	Var.	Var.
Teacher	0.05	0.10	0.03	0.15	0.07	0.25
Residual	0.16	0.20	0.12	0.17	0.15	0.32
SD of teacher effect	0.22	0.32	0.16	0.39	0.26	0.50
SEM of a single observation	0.40	0.44	0.34	0.41	0.39	0.57
Reliability of a single observation						
Relative standing of teachers	0.24	0.35	0.18	0.48	0.30	0.43
Absolute scores of teachers	0.24	0.35	0.18	0.48	0.30	0.43
N (teachers)	156	156	156	156	156	156

Notes: The table shows variance components from a G-study of student-teacher interaction scores in the full sample. The object of measurement is the teacher and the facet of error is occasion, defined as one of the first ten scored student-teacher interactions in the lesson. The SD of the teacher effect is the square root of teacher variance. The standard error of measurement (SEM) of a single observation is the square root of relative error variance. Variance components estimated as negative are set to zero.

Table C.6: Variance components of student-teacher interaction scores, two-rater sub-sample (endline)

	(1) Initiator	(2) Direction	(3) Content	(4) Tone	(5) Framing	(6) Length
Variance component	Var.	Var.	Var.	Var.	Var.	Var.
Teacher	0.04	0.04	0.01	0.08	0.07	0.22
Rater	0.00	0.00	0.00	0.00	0.00	0.00
Occasion × teacher	0.15	0.16	0.13	0.08	0.11	0.28
Rater × teacher	0.00	0.01	0.01	0.00	0.00	0.00
Residual	0.04	0.05	0.02	0.07	0.03	0.06
SD of teacher effect	0.21	0.19	0.11	0.28	0.27	0.47
SEM of a single observation	0.13	0.16	0.11	0.20	0.13	0.18
Reliability of a single observation						
Relative standing of teachers	0.70	0.58	0.52	0.67	0.81	0.87
Absolute scores of teachers	0.20	0.16	0.08	0.41	0.36	0.41
N (teachers)	48	48	48	48	48	48

Notes: The table shows variance components from a G-study of student-teacher interaction scores in the random two-rater sub-sample. The object of measurement is the teacher and the facets of error are rater and occasion, defined as one of the first ten scored student-teacher interactions in the lesson. The SD of the teacher effect is the square root of teacher variance. The standard error of measurement (SEM) of a single observation is the square root of relative error variance. Variance components estimated as negative are set to zero.