Improving Math and Science Skills in Contexts of Low Teacher Capacity: Experimental Evidence from India*

Alejandro J. Ganimian† Harvard University

Isaac M. Mbiti‡ University of Virgina

Abhilash Mishra§ Equitech Futures/ University of Chicago

September 13, 2024

Abstract

We present experimental evidence on a program in India that recruited college students ("fellows") in math and science fields to teach in primary schools for a year. Fellows were younger, less educated, and less experienced than teachers, but they outperformed them by 1.4σ on a test of content knowledge and pedagogy. They received a brief training, lesson scripts, and instructional coaches. During unannounced visits, fellows were no more likely to go to work or arrive early than control teachers. Yet, during announced observations, they fared 0.73σ better on an index of positive instructional practices. After a year, their students scored 0.34σ higher in math, 0.22σ in science, and 0.15σ in language than those taught by regular teachers. By the start of the next year, they still scored 0.36σ, 0.14σ, and 0.08σ higher in these subjects, respectively. They did not, however, differ on attitudes towards math and science, intelligence and math mindsets, or aspirations to pursue related careers. Evidence of compensatory behavior (more resources) in control classes suggests our estimates may understate true effects.

JEL codes: C93, I21, I22, I25

Keywords: math and science education, teacher capacity, lesson scripting, alternative pathways into teaching, India.

*We gratefully acknowledge the funding provided by the Abdul Latif Jameel Poverty Action Lab's (J-PAL) Post-Primary Education initiative and the University of Chicago's Kevin Xu Initiative on Science, Technology, and Global Development for this study. We thank Priya Pethe and Rahul Panda for making this study possible. We also thank Austin Dempewolff, Sharnic Djaker, Mustufa Patel, and especially Rashmi Menon, who provided excellent research assistance. Finally, we thank seminar participants at New York University, the University of Virginia, Harvard University, and the London School of Economics and Political Science for useful comments. We registered this study with the AEA Trial Registry (RCT ID: AEARCTR-0002386). Research Protocols were approved by the Institutional Review Boards at UVA and the Institute for Financial Management and Research. All views expressed are our own and not of the institutions with which we are affiliated.

†Harvard. alejandro_ganimian@gse.harvard.edu.

‡UVA, NBER, J-PAL, and BREAD. imbiti@virginia.edu.

§Equitech Futures and University of Chicago. abhilash@equitechfutures.com.

1 Introduction

There is growing evidence on the individual and social returns to math and science skills. Across very different contexts, individuals with higher mathematical and scientific literacy complete more years of education, are more likely to be get a job, and earn more once employed. For example, in the United States, each standard deviation (SD) in math test scores at the end of secondary school is associated with 12% higher earnings (Mulligan, 1999; Lazear, 2003).[1] In India, those who study math and science in high school complete more years of schooling and have 22% higher earnings than those in either business or humanities (Jain et al., 2022).[2] Further, a country's scores on international assessments of math and science predict its level of innovation and economic growth (Bianchi and Giorcelli, 2020; Hanushek and Kimko, 2000; Barro, 2001; Woessmann, 2003; Jamison et al., 2007; Hanushek and Woessmann, 2008).

Teachers play a pivotal role in the production of the math and skills of their students. In Peru, each SD on a test of teachers' content knowledge increased student achievement by about 0.1 SDs (Metzler and Woessmann, 2012).[3] Yet, their capacity on these subjects is low. Secondary-school students hoping to become teachers score below the national mean in math on global tests, and a fourth to a full SD below aspiring engineers (Bruns and Luque, 2014).[4] The traditional approach to this challenge, professional development, has been ineffective (Loyalka et al., 2019; Popova et al., 2022), possibly because it seeks to improve pedagogy without addressing gaps in teachers' content knowledge. If the two are complements in the education-production function, focusing on one of them will not be enough (Mbiti et al., 2019).

We evaluated a novel intervention in India that sought to improve math and science skills by both shifting the composition of teachers (to raise their content knowledge) and the supports they received (to improve their pedagogy). The Science Education Initiative (SEI) recruited college students ("fellows") majoring in math and science to teach for a year in primary schools. Fellows were selected through a competitive process and received pre- and in-serving training, lesson scripts, and periodic feedback from instructional coaches. They taught for six hours per week and received a small stipend of about USD 4 per lesson. We conducted a randomized evaluation of this intervention in Pune, the second-largest city in the state of Maharashtra. We randomly assigned 48 public and charter schools to receive the program in grade 5 or 6.[5]

[1] See also Bishop (1989); O'Neill (1990); Grogger and Eide (1995); Murnane et al. (1995); Neal and Johnson (1996); Mulligan (1999); Altonji and Pierret (2001); Murnane et al. (2001) or Hanushek (2002) for a review.

[2] See also Boissiere et al. (1985); Alderman et al. (1996); Glewwe (1996); Angrist and Lavy (1997); Jolliffe (1998); Molina et al. (2020); Behrman et al. (2008) or Hanushek and Woessmann (2008) for a review.

[3] This finding is consistent with correlational evidence across 31 countries (Hanushek et al., 2019)

[4] Gaps in teachers' content and pedagogical knowledge are even more pronounced in LMICs. For example, a study in Sub-Saharan Africa found two thirds of math teachers could not solve an algebra problem and nine in ten could not prepare a lesson using a text (Bold et al., 2017).

[5] The 26 grade 5 and 25 grade 6 classes that received the intervention made up the treatment group and the rest served as controls. This randomization strategy, pioneered by Banerjee et al. (2007), ensures that

Relative to incumbent teachers, fellows were younger and more likely to be male, they had less teaching experience, and they had completed fewer years of formal education.[6] Despite their lower levels of education and lack of teaching experience, fellows scored 1.4σ above regular teachers in a written assessment of math and science knowledge, instructional practices, and understanding of students' misconceptions. Among those found in school, fellows were more likely than control teachers to be in the classroom (56% v. 4.8%, $p < 0.01$), as opposed to somewhere else in the school. Yet, despite differences in classroom presence, classroom observations (which are conditional on classroom presence) show little differences in student engagement and other measures of lesson-time allocation between fellows and teachers. However, they do reveal a clear substitution pattern—in classrooms with fellows, regular teachers reduce their instructional effort, leaving the science and math lessons to the fellow, instead of collaborating with them. Thus, the intervention essentially replaces existing teachers with less trained and less experienced fellows who have more subject-matter knowledge and scaffolding support (scripted lessons plus coaching).

We present three main findings. First, the intervention had moderate-to-large positive impacts on student achievement. After accounting for baseline test scores, treatment students outperformed their control peers by 0.34 standard deviations (SDs) in math and 0.22 SDs in science ($p < 0.01$ for both).[7] During the study, we found that some items from the baseline test were used to coach students, but test-score impacts cannot be explained by such coaching. Treatment students still outperformed their control peers on items introduced at endline and follow-up.[8] They also fared better on contemporaneous tests drawing from a different source of items, although the material in these tests may have been better aligned with instruction in science.[9] Further, they outperformed control counterparts at the start of the next school year by 0.36 SDs in math ($p < 0.01$) and 0.14 SDs in science ($p < 0.05$). Both the end-of-year and next-year gains in student achievement were broad-based, with the treatment distributions stochastically dominating the control distributions for both rounds.[10]

Second, even though the fellows only taught math and science, student achievement in language also increased by 0.15 SDs ($p < 0.01$), suggesting that the improvements in math

all schools have an incentive to participate in all data-collection rounds. We discuss it in greater detail in section 2.3.

[6]Due to frequent transfers across schools, regular teachers on average only had about 2 years teaching math or science at their school, so the intervention did not sharply reduce the teaching experience of the average instructor in these subjects.

[7]Scaling results to account for differences in item characteristics yielded nearly identical results. Impacts are equivalent to 6.7 pp. in math and 3.8 pp. in science.

[8]After accounting for baseline, treatment students outperformed controls in non-repeated items at endline by 4.1 pp. in math ($p < 0.01$) and 1.7 pp. in science ($p < 0.05$), and at follow-up by 4.1 pp. in math ($p < 0.01$) and 1.5 pp. in science ($p < 0.05$).

[9]After accounting for baseline, treatment students outperformed control peers by 0.09 SDs in math (not significant) and 0.40 SDs in science ($p < 0.01$).

[10]We found no heterogeneous effects by students' sex or caste. We found higher impacts for students with higher socio-economic status, but they mostly become statistically insignificant once we account for baseline.

and science did not come at the expense of other subjects. This positive spillover is consistent with the perceived relationship between science literacy and foundational literary (**?**), through which activities such as reading science texts and expanding students' (scientific) vocabulary could improve broader literacy skills. It could also reflect the increased ability of teachers to focus on other subjects because they are responsible for fewer lessons in science and math.

Third, despite these learning gains, and the potential for fellows to serve as role models, the intervention did not change students' attitudes towards math and science, growth mindset (specifically relating to intelligence and math), or aspirations to pursue related careers.[11] The null effects on these indicators suggest that the intervention did not raise achievement by boosting students' motivation or their perceived payoff from exerting effort in these subjects.

The intervention features a bundle of complementary inputs, so we cannot distinguish between the effects of each component. Yet, we find some evidence consistent with lesson scripts playing an important role. If fellows adhered to such scripts, we would expect similar treatment effects across all sites. This is what we find: we fail to reject the test that the effect of the intervention was the same across all schools, using the joint test of homogeneity developed by von Hippel and Bellows (2018). Further, if scripts are driving impacts, the importance of instructors' content knowledge should be dampened when teaching with such scripts script. Again, this is what we find: there is a negative interaction effect between instructor's content knowledge and the intervention, even if there is correlation between such knowledge and students' achievement. Taken together, these results suggest imperfect script adherence, which is consistent with the data collected from our classroom observations.

We also present three sets of robustness checks. We first show that the intervention did not increase students' attendance to school, ruling out the possibility that it improved achievement by motivating students to go to school more regularly. We then show that it did not increase students' propensity to attend private tuition in math and science or time spent on tuition, ruling out students seeking outside help on these subjects as a primary mechanism of the impact. We also find that control classes were more likely to have learning and teaching materials than treatment classes, suggesting that principals may have tried to compensate the former. If correct, our estimates would actually understate the true impact of the intervention.

Our first contribution is to global evidence on what teachers know and are able to do. Internationally comparable data have traditionally tracked teachers' education and experience (OECD, 2022; UNESCO, 2022; World Bank, 2023), but such metrics do a poor job predicting teaching effectiveness (Rockoff, 2004; Rivkin et al., 2005; Kane et al., 2008; Kane and Staiger, 2008; Rockoff et al., 2011; Kane and Staiger, 2012; Kane et al., 2013; Araujo et al., 2016).

[11] Most students enjoyed studying these subjects, but expressed some performance anxiety. About half believed intelligence and math skills cannot be developed, and almost as many thought boys were smarter and better at math than girls. Three-fourths wanted to study math or science in high school, more than two-thirds wanted to continue studying after high school, and a third aspired to a job involving math or science.

Studies in LMICs have drawn attention to the high teacher absence rates in these settings (Kremer et al., 2005; Chaudhury et al., 2006; Muralidharan et al., 2017), and more recently, to how lesson time is allocated (Abadzi, 2009; Bruns and Luque, 2014; Stallings et al., 2014). Yet, we still have a relatively narrow understanding of teachers' competence in these contexts.[12] Our study presents a rare comprehensive account of teachers' work in a developing setting, drawing on surveys to describe their background and credentials, assessments to measure their content and pedagogical knowledge, surprise school visits to track their attendance, and class observations to monitor time allocation and identify the prevalence of instructional practices.[13]

Grounded in this understanding, we also advance causal research on how to improve math and science instruction in contexts of low teacher capacity. Prior work largely takes gaps in the preparation of the current stock of teachers as given and focuses on mitigating its effects, automating parts of their work—e.g., using pre-recorded lessons (Naslund-Hadley et al., 2014) and lesson segments (Beg et al., 2022; de Barros, 2022)—or asking students to learn on their own—e.g., with inquiry-based (Beuermann et al., 2013; Bando et al., 2019) and peer-to-peer learning (Wachanga and Mwangi, 2004; Ajaja and Eravwoke, 2010; Berlinski and Busso, 2017). Our study shows that it is possible to raise the capacity of the teaching labor force by recruiting individuals with subject-matter expertise and providing them with pedagogical support.[14]

We also contribute to the literature on scripted lessons in LMICs. Prior studies have shown that, in contexts of extremely low teacher capacity (e.g., little or no post-secondary education), scripts that tell teachers what to do at each stage of a lesson can ensure a minimum "floor" of instructional quality (Piper and Korda, 2010; Tilson et al., 2013; Piper et al., 2018; Albornoz et al., 2020; Romero et al., 2020; Gray-Lobe et al., 2022; Eble et al., 2021; Lucas et al., 2014). In fact, structured pedagogy more broadly was recently identified as one of the most cost-effective interventions to improve learning outcomes in LMICs (Akyeampong et al., 2023). Our study illustrates how such scripts can also be provide scaffolding to instructors who have high levels of subject-matter expertise but low levels of pedagogical training and experience. It is also one of the first to document the extent to which instructors adhere to the different components of scripts and to examine which types of teachers are more likely to use them.

The rest of the paper is structured as follows. Section 2 describes the context, intervention, sampling, and randomization. Section 3 presents the data. Section 4 discusses the empirical strategy. Section 5 reports the results. Section 8 discusses implications for policy and research.

[12]International surveys of teachers (e.g., the International Evaluation Association's Teacher Education and Development Study in Mathematics [TEDS-M] or the Organization for Economic Cooperation and Development's Teaching and Learning International Survey [TALIS]) offer rich descriptions of teachers' work, but only a handful of LMICs have ever participated in them (Tatto et al., 2012; OECD, 2019).

[13]For similar work, see Bhattacharjea et al. (2011); Bold et al. (2017); World Bank (2017).

[14]To our knowledge, there is only one other study on a similar approach in a LMIC, but it is not causal (it uses propensity-score matching), it was conducted in a country subsequently categorized as high income (Chile), and it evaluates an initiative that does not intend to keep recruits in teaching (Alfonso et al., 2010).

2 Experiment

2.1 Context

We conducted our study in the city of Pune, the second-largest city in the Indian state of Maharashtra (after its capital city, Mumbai). According to the latest census of India (in 2011), there are 9.4 million people in Pune (NIEPA, 2017), rendering its population size comparable to that of countries such as Belarus, Honduras, and the United Arab Emirates (UN, 2019). Its school system is run by the Pune School Board (*Shikshan Mandal* or PSB) under the Pune Municipal Corporation (PMC). On the latest school year with available data, there were 3,473 primary-only schools (grades 1-5) and 1,941 additional primary schools with upper primary (grades 6-8), with 834,354 students enrolled in both types of primary schools (NIEPA, 2017). For reference, if Pune were a school system in the U.S., it would rank between the top two largest districts in number of students: New York City and Los Angeles Unified (NCES, 2018).

Nearly all primary-school aged children in Pune are enrolled in school: the gross enrollment rates are 110% in primary and 109% in upper primary.[15] Marathi is the language of instruction in 76% of primary-only schools and 55% of those with upper primary; English-medium schools account for 23% and 41% of these types of schools, respectively.[16] About 85% of primary-only schools and 63% of primary schools with upper primary are "government" (i.e., public) schools. There are, on average, 24 students per teacher in primary-only schools and 32 in primary schools with upper primary, and these numbers closely track class sizes (NIEPA, 2017).

The public sector employs most primary-school teachers: 61% of teachers in primary-only and 47% of those in schools with upper primary work in government schools. The vast majority of teachers in these types of schools are "regular" (i.e., tenured): 93% and 84%, respectively; the rest are hired on a renewable contract basis (NIEPA, 2017).

Most primary-school children in Pune lack basic math and science skills. According to a nationally representative student assessment administered by the central government, only 30% of fifth-graders in Pune could use arithmetic for daily situations, 38% could identify equivalent fractions, and 34% could estimate a volume. Results for science were equally discouraging: 23% could identify linkages between terrain, climate, and resources; 29% could group objects, materials, and activities according to properties such as shape, color, and sound; and 45% could estimate spatial quantities in simple standard units (NCERT, 2018).

[15] The gross enrollment rate indicates the number of children enrolled at a given education level *irrespective of age*, divided by the number of children *of age* to attend this level and multiplied by 100. Gross enrollment rates often exceed 100% because the denominator includes both younger and older children. The net enrollment rate (the number of children enrolled at a given level who are of age to attend such level, divided by the total number of children of age for that level) is only reported for upper primary and it is 91% (NIEPA, 2017).

[16] According to anecdotal evidence, a non-trivial share of instruction in these schools is also in Marathi.

2.2 Sample

We selected 48 public and *Vidya Niketan* primary schools for this study.[17] We started with all 286 PSB-run primary schools. We excluded 118 schools away from the city center (because their location would have limited the capacity of the non-profit running the intervention to monitor its implementation), 30 Urdu-medium schools (because most fellows did not speak Urdu),[18] 59 schools where the PSB or other non-profits were conducting other programs (because we wanted to estimate of the effects of the intervention on its own), 20 schools with low enrollment (to minimize sampling error), and nine schools that already participated in the intervention (because we wanted to estimate the effects of the first year of the intervention). Our data-analytic sample includes 46 of the 48 sampled schools. Shortly after baseline, we had to drop two schools that could not be matched to any fellows based on their preferences.

2.3 Randomization

We randomly assigned the 48 sampled schools to receive the intervention in grades 5 or 6. This process resulted in 26 grade 5 and 25 grade 6 treatment classrooms and 26 grade 5 and 25 grade 6 control classrooms, such that all schools had at least one classroom with a fellow.[19] This randomization strategy, pioneered by Banerjee et al. (2007), seeks to minimize the risk of differential attrition (i.e., schools without any intervention dropping before the endline).

We also randomly assigned fellows to schools, conditional on their preference set. First, we grouped fellows based on their preferred neighborhood, school shift (morning or afternoon), and medium of instruction (English or Marathi). Then, we ran 17 lotteries—one per preference set (e.g., one lottery for neighborhood 1, morning shift, English medium schools).

Table 1 presents summary statistics on students and compares the characteristics and achievement of students between experimental groups. The mean control-group student was 11 years old, which is expected given that the sample is split between grades 5 and 6. Most control students (69%) speak Marathi at home, some (17%) speak Hindi, and few (1%) speak English. Less than two-thirds of them have mothers who completed primary school and more than three-fourths have fathers who reached this level. Nearly all of them (90%) have a TV, but fewer have Internet (43%), a desk (28%), a computer (20%), or their own room (17%),

[17] These schools are publicly funded and managed like regular public schools, but they have more autonomy over school-management decisions.

[18] Note, however, that Urdu-medium schools account for about 1% of primary-only schools and 3% of schools with upper primary in Pune (NIEPA, 2017).

[19] Two schools had two grade 5 and two grade 6 classrooms and two other schools had *either* two grade 5 *or* two grade 6 classrooms. In both cases, we assigned all classrooms within the same grade to the same experimental group to prevent contamination, which is particularly likely to occur within the same grade (e.g., a grade 5 fellow in a treatment classroom sharing materials with the grade 5 teacher in a control classroom). All other schools had one treatment and one control classroom (either grade 5 or grade 6).

suggesting that the schools where SEI places its fellows serve relatively low-income families. Yet, two in three of these students attends tuition in math and one in four does so in science.[20]

We find no systematic differences in the characteristics or achievement of students across experimental groups. By chance, treatment students had lower achievement in math ($p < 0.1$), so we estimate the effect of the program with and without accounting for baseline achievement.

2.4 Intervention

The program that we evaluated was the Science Education Initiative's Fellowship Program.[21] SEI is a Pune-based non-profit organization dedicated to improving math and science learning and the fellowship was its flagship program. Since 2014, it has placed 200 fellows in 110 classes. It recruits, trains, and supports college majors ("fellows") in science, technology, engineering, and math to teach math and science in schools serving disadvantaged students for one year.

The intervention has four main components. The first one is changing the composition of math and science instructors. Fellows are selected through a four-stage process, including: (a) a test of content knowledge based on the math and science curricula for grades 5 to 10;[22] (b) a brief (5- to 7-minute) demonstration lesson on a topic chosen by each applicant; (c) an interview to assess applicants' scientific aptitude, leadership skills, and motivation to teach; and (d) a longer (15- to 20-minute) demonstration lesson on a topic chosen by SEI.[23]

Fellows differ from teachers in background, education, experience, and knowledge (Table 2). They are less likely to be female (63% v. 76%) and nearly half their age (21 v. 42 years old). Only one in three fellows has a bachelor's degree (most are still completing it), compared to eight in ten teachers. By the end of the study, the average fellow had completed their first year of teaching, whereas the average teacher had accrued 18 years of teaching experience, five of which had been at their current school, and two of which focused on math or science. Yet, despite their lower education and experience, fellows outperformed teachers by 14 pp. on a test of content knowledge, instructional practices, and student misconceptions.[24]

The second component of the intervention is providing pre- and in-service teacher training. Before they start teaching, fellows must complete a three-week course on pedagogy, classroom management, and the math and science curriculum. Once they begin the fellowship, they are also enrolled in a bachelor's in education paid for by SEI at a local teacher-training college.

[20] Private tuition is common in urban India, even among low-income families (Berry and Mukherjee, 2019).

[21] After our study, SEI changed its name to Science for All.

[22] These curricula are jointly determined by the National Council of Educational Research and Training and the Board of Education of Maharashtra.

[23] There are no language requirements, but applicants proficient in Marathi are prioritized, given that most primary schools in Pune are Marathi-medium schools.

[24] We describe this test in section 3. Fellows also vary less in their scores (see Figure A.1 in appendix A).

The third component is scripting instruction. Fellows are given lesson scripts for every lesson, which specify the topics to be taught on each day, the materials to be used, the words to be written in the blackboard, the activities to complete, and the questions to ask students.

Most fellows wrote on the board and asked students questions as specified on the script, but only half used the suggested materials and activities, and a fifth to a third made changes (adding to, rephrasing, or excluding parts of the lesson script; see Table A.1 in appendix A).[25] Interestingly, scripts served as a *complement* to subject-matter expertise: fellows who scored above the median in the test of content knowledge, instructional practices, and understanding of students' misconceptions mentioned above were *more* likely to follow the scripts.

The fourth component is substituting part of the lesson time otherwise assigned to teachers. Schools in Pune do not have timetables that specify the number of minutes each lesson is supposed to last or even the number of lessons assigned to each subject, so we cannot calculate the share of lesson time in math and science for which fellows substituted regular teachers. Yet, we know fellows were expected to teach three times per week, and that on each occasion, they taught both math and science for 120 minutes (i.e., 24 hours per month) for one year. As compensation, fellows received a stipend of INR 3,000 (USD 47) per month.[26]

Our study design allows us to estimate the combined effect of the three components above. We believe this is the parameter of policy interest, as it would make little sense to place college students inside classrooms and expect them to teach well without any training or supports. We also suspect that it is in part this training and supports that lead many college majors in math and science to apply to the program (instead of trying to enter teaching on their own), so removing them could have knock-on effects on the composition of the pool of applicants, rendering an evaluation of the change in the composition of instructors by itself challenging. Lastly, the training and supports do not account for a large share of the program's expenses.[27]

3 Data

As Table 3 shows, we conducted four rounds of data collection, including: (a) student surveys and assessments at baseline (to check the comparability of experimental groups, increase the precision of our estimates, and test for heterogeneous effects); (b) unannounced school visits (to estimate the impact of the intervention on instructor attendance and punctuality) and announced classroom observations (to estimate effects on lesson-time allocation and pedagogical practices) during the school year; (c) student surveys and assessments (to

[25] We describe the announced classroom observations on which these figures are based in section 3.

[26] At the time of the study, the median teacher salary in Pune was about INR 50,000 (USD 601) per month.

[27] The total cost of running the fellowship in 2017-2018 was INR 3,662,000 (USD 56,867). With 60 fellows and 1,920 students that year, it cost INR 61,033 (USD 948) per fellow and INR 1,907 (USD 30) per student. Of those 60 fellows, only 48 are part of the present study.

estimate impacts on achievement, attitudes, mindsets, and aspirations) and instructor surveys and assessments (to compare PMC teachers with SEI fellows) at endline; and (d) student assessments at follow-up (to check whether impacts on achievement persisted over time).

3.1 Student surveys

We administered a short student survey at baseline focusing on background characteristics (e.g., sex and socio-economic status) to test for heterogeneous effects, and a longer one at endline measuring constructs that may be affected by the intervention (e.g., attitudes towards math and science, intelligence and math mindsets, and educational and career aspirations).

3.2 Student assessments

We administered student assessments of math and science (the two subjects targeted by the intervention) and language (to test for spillover effects) at baseline, endline, and follow-up. All assessments evaluated what students know and are able to do based on global standards.[28] Each test had 30 multiple-choice items.[29] We included items from a wide range of difficulty levels to reduce the possibility of "floor" and "ceiling" effects (i.e., students answering no or all questions correctly, respectively).[30] We present the results in three ways: percentage-correct scores (i.e., the percentage of items answered correctly), standardized scores (at baseline, with respect to the overall distribution; at endline and follow-up, with respect to the control group), and Item-Response Theory (IRT) scores to account for differences in students' ability and item characteristics (i.e., difficulty and differentiation between students of similar ability).[31]

Right before the endline, we noticed that SEI's instructional coaches had used items from the baseline assessments in practice tests. We had asked the organization not to read or keep assessments, but internal miscommunication resulted in coaches making copies of the tests. This was problematic because it raised the possibility that any effects observed at endline may be solely due to coaching on repeated items (which are needed to link results across rounds). We test for this possibility in three ways. First, we estimate endline effects separately for repeated items (which were first administered at baseline and were thus subject to coaching)

[28]The math and science tests were based on the 2019 Trends in International Math and Science Study assessment framework (IEA, 2017). The math test covered three content domains (numbers, geometry and measurement, and data display) and three cognitive domains (knowing, applying, and reasoning). The science test covered three content domains (life, physical, and earth science) and the same cognitive domains as the math test. The language test was based on the 2016 Program for International Reading Study framework (IEA, 2015). It covered three content domains (vocabulary, grammar, and reading) and three cognitive domains (retrieving explicit information, making inferences, and interpreting and integrating ideas and information).

[29]At baseline and follow-up, we created two versions of each assessment to prevent cheating.

[30]We used items from international assessments, domestic assessments, and impact evaluations in India. Figures A.2 and A.3 shows the distributions of proportion-correct and IRT-scaled scores, respectively.

[31]We used a two-parameter logistic IRT model (Yen and Fitzpatrick, 2006).

and non-repeated items (which were introduced at endline and thus not subject to coaching). Then, we estimate endline effects on "audit" assessments of math and science, which drew on different items from a concurrent evaluation (Gray-Lobe et al., 2022), to a subset of students. Lastly, we also estimate effects on follow-up assessments of math, science, and language at the start of the next school year to all students in our sample.

3.3 Unannounced school visits

We conducted unannounced visits to school during the school year to estimate the impact of the intervention on instructor attendance and punctuality by comparing SEI fellows to PMC teachers in control and treatment classes. We did not announce these visits to minimize the chances that instructors would attend school, or would do so earlier than usual, because of us. We also collected administrative data and counted the number of students in the classroom to estimate the impact of the intervention on student attendance and punctuality.[32]

3.4 Announced classroom observations

We conducted announced classroom observations during the school year to estimate the impact of the intervention on instructor lesson-time allocation by comparing SEI fellows to PMC teachers in control and treatment classes.[33] We announced our observations because we were interested in how teachers used lesson time when they attended.[34] We also collected data on whether fellows and control teachers engaged in certain practices during the lesson.

3.5 Instructor surveys

We administered a survey of instructors at endline to compare SEI fellows to PMC teachers on background characteristics (e.g., sex, education, training, and experience) to understand how the intervention had changed the composition of instructors to which students were exposed.

3.6 Instructor assessments

We administered instructor assessments at endline to compare SEI fellows to PMC teachers on content knowledge, instructional practices, and understanding of students' misconceptions.

[32] We supplemented these measures with survey-based measures from endline.

[33] We adapted a classroom-observation protocol that has been widely used in LMICs, including India (Stallings, 1977; Bruns and Luque, 2014; Sankar and Linden, 2014; Stallings et al., 2014; World Bank, 2017).

[34] As stated above, schools in Pune do not have timetables. However, based on conversations with principals, we observed PMC teachers for 30 minutes per subject, which seemed to be the modal duration of a lesson. We observed fellows for their entire 120-minute slot, during which they taught both math and science.

The test had 36 multiple-choice items.[35] The items on content knowledge were sampled from the student assessments. Those on instructional practices presented objectives for hypothetical lessons and asked respondents to choose their preferred approach to pursue those goals. Those on student misconceptions presented mistakes that students made and asked respondents to identify the most likely underlying reason for students' misunderstanding.

4 Empirical strategy

We estimate the intent-to-treat effect of the offer of the intervention by fitting the model:

$$Y_{igs}^{t} = \alpha_{r(g)} + Y_{igs}^{t=0}\gamma + T_g'\beta + \epsilon_{igs} \tag{1}$$

where Y_{igs}^{t} is the outcome of interest for student i in grade g and school s at endline ($t = 1$) or follow-up ($t = 2$); $Y_{igs}^{t=0}$ is a measure of that outcome at baseline (when available); $r(g)$ is the randomization stratum of grade g and $\alpha_{r(g)}$ is the corresponding stratum fixed effect; T_g is an indicator variable for random assignment to the intervention; and ϵ_{igs} is an error term. The parameter of interest is β, which captures the causal effect of the offer of the intervention. We estimate equation (1) by ordinary least-squares regression. We use cluster-robust standard errors to account for within-school correlations across students in outcomes.

We also fit variations of this model in which outcomes are measured at the instructor level (to estimate the impact of the intervention on instructors) and models that interact the intervention indicators with student and teacher covariates (to test for heterogeneous effects).

5 Results

5.1 Student achievement

After one year, the intervention had moderate-to-large effects on math and science test scores. As Table 4 shows, the endline scores of treatment students were 0.247 SDs higher in math and 0.207 SDs higher in science than those of their control peers ($p < 0.01$). If we account for baseline scores, estimates are slightly higher (0.340 and 0.216 SDs, respectively, $p < 0.01$).[36] We observe improvements in all three content domains (numbers, geometry and measurement, and data display) and two of the three cognitive domains (knowing and applying) assessed

[35] We drew on items from international assessments, domestic assessments, and previous assessments of teacher knowledge and skills conducted by domestic organizations (e.g., Pratham, Educational Initiatives) and international organizations (e.g., Educational Testing Service, Bridge International Academies).

[36] These effects are equivalent to 6.7 pp. in percent-correct scores in math and 3.8 pp. in science (Table A.6). We obtain very similar results if we scale the endline test scores using a two-parameter logistic IRT model to account for differences in students' latent ability and item characteristics (Table A.7).

in math, and all content domains (life, earth, and physical science) and cognitive domains in science (knowing, applying, and reasoning).

These effects are not entirely explained by fellows coaching students on specific items.[37] As Table 5 shows, effects on items that first appeared at baseline are larger (7.1 pp. in math and 5.4 pp. in science, respectively, $p < 0.01$) than those on those that we introduced at endline (2.8 pp. in math and 1.8 pp. in science, $p < 0.01$ and $p < 0.05$, respectively), and for most specifications, we can reject the null that effects on these two sets of items are equal.[38] Yet, the impacts on the latter are statistically significant, suggesting actual learning gains.

A parallel test that we administered at endline offers further evidence that coaching does not fully account for the impacts.[39] As panel B of Table 4 shows, treatment students also outperformed control students in this "audit" test by 0.046 SDs in math and 0.367 SDs in science. Only the effects on science are statistically significant ($p < 0.01$), but this may be because the test may have overlapped more with the material taught in that subject.

The fact that treatment students still fared better than their control peers at the start of the following school year confirms that the intervention impacted learning beyond coaching. As panel C of Table 4 shows, treatment students scored 0.283 SDs higher than control students in math and 0.120 SDs in science or 0.356 SDs and 0.142 SDs accounting for baseline scores (in both specifications, $p < 0.01$ and $p < 0.05$, respectively).

The intervention also had spillover effects on language. As Table 4 indicates, the treatment group outperformed the control group by 0.132 SDs or 0.151 SDs once we account for baseline scores ($p < 0.05$ and $p < 0.01$, respectively), even if fellows did not teach this subject.[40] We see effects in all three content domains (vocabulary, grammar, and reading) and two of the three cognitive domains (retrieving information and making inferences) assessed in language. In fact, treatment students still outperformed their control counterparts in language at the start of the following school year by 0.059 SDs, but the effect is only statistically significant if we account for baseline (0.081 SDs, $p < 0.05$).

The intervention improved test scores across all levels of the achievement distribution. Quantile treatment effect plots show that the treatment distributions first-order stochastically dominate the control distributions on the endline, audit, and follow-up tests, suggesting that the intervention led to broad-based gains (Figure A.4). Non-parametric estimates of average treatment effects at each percentile of the baseline scores also show large positive impacts for

[37] As we discuss in section 3.2, right before the endline, we noticed that the nonprofit running the program had included some of our baseline items in its materials.

[38] Remarkably, the differences between effects on repeated and non-repeated items in math and science remain stable at followup, suggesting the effects of coaching persisted (Table 6).

[39] As we mention in section 3.2, this test drew on an entirely different item bank and was used in a concurrent impact evaluation (Gray-Lobe et al., 2022).

[40] This effect translates into 3.5 pp. (Table A.6). IRT-scaled impacts are very close (Table A.7). Effects on language, however, are no longer statistically significant at follow-up.

all three rounds of data collection across the full range of baseline achievement (Figure A.5). Consistent with these results, we find no heterogeneous effects by baseline scores (Table A.8).

We find some evidence that the intervention was more effective for students with higher socio-economic status. In our estimation of endline effects, the interaction term between the treatment and the first principal component of a principal-component analysis of home assets (SES index) is positive and statistically significant for all subjects, and the p-value of the sum of the main and interaction effects of this index is statistically significant for math and science (Table A.9, panel A, cols. 1, 3, and 5). Once we account for baseline performance, however, the p-value of the sum is no longer statistically significant for any subject (cols. 2, 4, and 6). Further, we do not see any evidence of heterogeneity by SES in the endline audit (panel B). The interaction between the treatment and SES index is positive and statistically significant for the follow-up (panel C, cols. 1, 3, and 5), but this interaction and the p-value of the sum only remain statistically significant for science after we control for baseline (cols. 2, 4, and 6).

We do not find evidence of heterogeneous effects by other student or teacher characteristics. The interaction between treatment and an indicator for female students is positive for all subjects at endline, but it is statistically significant only for math once we account for baseline (Table A.10, panel A) and statistically insignificant in the other two rounds (panels B and C). The interaction between treatment and an indicator for students from scheduled castes or tribes is negative but statistically insignificant for nearly all subjects and rounds (Table A.11). Lastly, students do not benefit more from the intervention when they have an instructor of the same sex (Table A.12) or one who scores higher on the written assessment (Table A.13).[41]

6 Mechanisms

6.1 Instructor pedagogical practices

The intervention seems to have impacted test scores mainly by improving the pedagogical practices instructors used—specifically, by increasing the frequency of positive practices. As panel B of Table 7 shows, control teachers used closed- and open-ended questions (63%), asked students to explain their answers (47%), corrected wrong answers (72%), allowed students to ask questions (39%), provided individual help (68%), assigned homework (48%), and praised or encouraged students (76%). Fellows were even more prone to pursue these practices. In fact, they scored 0.73 SDs higher on a composite index of all positive practices ($p < 0.01$).

The introduction of fellows did not affect the frequency of negative practices, which were already rare. As panel A shows, only 19% of control teachers taught from the same spot and

[41] In fact, if we estimate the effect of the program only among high-scoring instructors (i.e., PMC teachers and SEI fellows), the point estimates (Table A.14) resemble the average effects of the intervention (Table 4), suggesting that effects are not explained solely by a selection effect.

5% or fewer remained sitting down, used their phone, got upset at incorrect answers, or was aggressive towards students. Fellows were 13 pp. less likely than control teachers to teach from the same spot ($p < 0.05$), but they were also 12 pp. more likely to get upset at incorrect answers ($p < 0.1$), and they did not differ from control teachers on a composite index of all negative practices, or reduce treatment teachers' engagement in these practices.

7 Robustness checks

7.1 Instructor attendance, punctuality, and lesson-time allocation

The intervention did not raise students' achievement simply by increasing the frequency with which they saw their instructors. Fellows were no more likely than control teachers to go to work or arrive on time—partly, because the latter were already doing so at fairly high rates. As panel A of Table 8 shows, 84% of control teachers were present during unannounced visits, compared to 72% of fellows, and this difference was not statistically significant. In fact, fellows were 30 pp. *less* likely to arrive on time than control teachers ($p < 0.01$).[42]

The introduction of fellows did not increase the attendance or punctuality of teachers in treatment classes either. Treatment teachers were slightly more likely than their control counterparts to be present (88% v. 84%) and less likely to arrive on time (71% v. 74%) during the unannounced visits, but neither difference was statistically significant.

Conditional on being at school, however, fellows were far more likely to be in their class. As panel B shows, only 4.8% of control teachers were in their classroom, compared to 56% of fellows, a difference of almost 51 pp. ($p < 0.01$). Fellows did not impact the likelihood of treatment teachers of being in their classroom, which was nearly identical (4.7%).

The intervention did not raise achievement merely by increasing the share of lesson time devoted to instruction either—largely, because control teachers were already teaching for most of their lessons. As panel A of Table 9 shows, the average control teacher devoted 78% of their lesson to instruction during announced classroom observations, compared to 75% for the average fellow, and the difference between these groups was not statistically significant.[43]

In fact, fellows allocated instructional time similarly to teachers in control classes. Control teachers spent most of their lesson time lecturing and explaining (34%), asking and answering questions (18%), and assigning students classwork (13%; Table A.2 in appendix A). The corresponding figures for fellows were nearly identical (34%, 15%, and 16%, respectively).

[42]Unlike teachers, fellows were only expected to be at school when teaching. Thus, their lower attendance rate may simply reflect that they were expected to be at school less frequently than teachers.

[43]As discussed in section 2.4, while PMC teachers teach either math or science in 30-minute lessons, SEI fellows teach both math and science combined in 120-minute lessons. Therefore, while we compare these groups focusing on the proportion of lesson time devoted to each type of activity, we also report the number of minutes devoted to each category in panel B of Table 9.

In fact, fellows also resembled teachers in their use of class management and off task time (Tables A.3-A.4), suggesting that time allocation was not a primary mechanism of impact.

Treatment teachers, however, rarely engaged with the fellows while they were teaching. The typical teacher in this group spent 8% of a fellow's lesson teaching, 21% managing the class, and 69% being off task—61 pp. more than the typical control peer ($p < 0.01$).

7.2 Student attitudes, mindsets, and aspirations

The intervention did not improve students' achievement by increasing their motivation to learn math and science. Roughly eight out of ten control students liked learning math and science and almost as many found them useful (Table A.15, panel A). It did not decrease students' negative feelings towards these subjects either. A fourth of control students felt nervous about them, a third wished they did not have to study these topics, and a fifth gave up when the material is difficult, but fellows did not change this pattern. In fact, treatment and control score perform similarly on a composite index of these measures.

Fixed mindsets were common among control students and fellows did not dispel them. About half of control students believed intelligence and math skills cannot be developed, and four in ten thought boys are both more intelligent and skilled at math. The intervention had a negative but statistically insignificant effect on all of these beliefs (Table A.15, panel B).

Lastly, the intervention did not make students more likely to want to stay in school, study math and science, or pursue a STEM-related job. Over two thirds of control students planned to pursue post-secondary education, three fourths wanted to study a STEM subject in high school, and a third aspired to a STEM-related job. The intervention had statistically insignificant effects on these metrics (Table A.15, panel C).[44]

Consistent with these results, we find no evidence that impacts on achievement are driven by role-model effects. As we would expect any potential role-model effects to be stronger when the fellow "looks like" the student, we test for heterogeneity in student test-scores by student-instructor sex match in Table A.12. However, we do not find any statistically significant effects on the interaction between treatment and sex match.

7.3 Student attendance

The intervention did not impact students' achievement by increasing their attendance to school. We measured student attendance in three different ways: by calculating the share of present students observed during unannounced visits, digitizing school attendance records,

[44]We did not find any evidence of heterogeneous effects on any of these sets of outcomes by students' sex, caste, socio-economic status, or baseline achievement (Tables A.16-A.19).

and asking students how often they were late to or missed school. None of these approaches suggest the intervention impacted student attendance (Table A.20).

7.4 Student demand for tuition

The intervention did not impact achievement by making students more likely to seek private tuition. Between a fifth and a third of students in our sample receive tuition in one subject. The intervention did not impact their propensity to seek tuition in a target (math and science) or non-target subject (English and Urdu) or the amount of tuition received (Table A.21).

7.5 Miscellaneous

Given the lower performance of teachers relative to fellows in our assessments, and the positive relationship between teacher content knowledge and student achievement, it is possible that the results reflect the presence of more knowledgeable instructors (i.e., fellows) rather than any other facet of SEI's program. To ameliorate this concern, we restrict our analysis to the set of teachers and fellows where the distribution of assessment scores overlap (see Figure A.1). This restriction essentially drops the lowest scoring PMC teachers (and their students) from the analysis. In Table A.14 we find that, across all specifications, the treatment effects are similar in magnitude to our full-sample estimates, suggesting that the results are not driven by differences in the teaching assessment.

One important feature of scripted instruction is that if it is implemented with high fidelity, it can reduce the heterogeneity in teaching quality across schools (Gray-Lobe et al., 2022). Additionally, if scripts are adhered to, an individual teacher's content knowledge will be less predictive of student outcomes. We test these implications in Figure A.6 and Table A.13.

Following von Hippel and Bellows (2018), we construct "caterpillar plots" of the school-specific (ITT) treatment effects and conduct a joint test against the null hypothesis of homogeneity. To ameliorate concerns about multiple hypothesis testing, we pool science and math scores into a composite test and the von Hippel and Bellows (2018) procedure computes both standard and Bonferroni confidence intervals to account for the number of school-specific effects. For the endline and the follow-up we fail to reject the null hypothesis of homogeneity, but we reject it for the audit. Across the three sets of student tests, we always fail to reject the null for science, but we reject the null in the math audit and the math follow-up (see Figures A.6-A.8). Overall, these results provide some suggestive evidence on the importance of scripted instruction, especially in science.

We examine the heterogeneity in student outcomes by teacher and fellow content score. In our empirical specification we use the assessment score of fellows in treatment classrooms, and the teacher's score in control classrooms to better reflect who the actual instructor was. The

results are reported in Table A.13. For the endline, teacher and fellow scores are associated with higher student test scores, although this effect is imprecisely estimated in three (out of six) specifications. Further, we also find that treatment dampens this effect, but this dampening is imprecisely estimated in five (out of six) specifications. While the patterns observed in the endline data are broadly consistent with the scripted instruction prediction, we do not find similar patterns in the endline audit or the follow-up.

Although these results provide some suggestive evidence consistent with scripting, especially in science, they also potentially shed light on the implementation fidelity. In Table A.1 we use data from announced classroom observations of fellows and document their script adherence on multiple dimensions. Overall, we find moderate levels of adherence. For instance, three-quarters of fellows asked students the questions in the script, about half of fellows used materials from the script, and a third added or expanded to the script. The patterns in Figure A.6 and Table A.13 are consistent with these moderate (to low) levels of script adherence.

7.6 Compensation for control classes

Whenever resources (e.g., fellows) are (randomly) allocated to some classes and not others, principal may try to compensate the non-selected classes with other resources (e.g., materials). To explore this possibility, we leveraged the information on class materials collected during the announced observations to compare the availability of materials across experimental groups. Control classrooms tend to have more materials than treatment classrooms, and they are statistically significantly more likely to have textbooks for teachers (52 pp.) and students (46 pp.; $p < 0.01$ in both cases) and math or science equipment (14 pp., $p < 0.05$; Table A.22). These results suggest that our estimates may understate the true impact of the intervention (i.e., the impact if all resources were equally distributed across control and treatment classes).

8 Conclusion

Developing math and science skills can help individuals improve their economic prospects. Yet, there is very little evidence on how to improve instructional quality in these subjects—especially, in contexts with low teacher capacity, which are more prevalent among LMICs.

We present experimental evidence on a novel approach to raise teacher capacity in math and science: hiring college students in these fields and providing them pedagogical support. We find that the introduction of these fellows results in moderate-to-large gains in student achievement in both of the subjects that they teach as well as in other subjects, and that those gains persist over time, influencing their students' preparation level for the next school year. We also identify potential mechanisms of impact (e.g., increase in instructors' likelihood to be

in their class, increase in their use of positive pedagogical practices) and rule out confounders (e.g., increase in instructor or student attendance or in student demand for private tuition).

These findings shed new light on the importance of teachers' subject-matter expertise. Until recently, its value had been largely inferred from correlations with students' performance on standardized tests (Hill et al., 2005; Santibanez, 2006; Rockoff et al., 2011; Kane et al., 2013; Gitomer et al., 2014; Bold et al., 2017; Cruz-Aguayo et al., 2017; Hanushek et al., 2019). We only know one quasi-experimental study on this question (Metzler and Woessmann, 2012). We provide experimental evidence that hiring instructors with high levels of content knowledge to replace teachers with lower levels for part of the school week raises student achievement.[45] The fellows in our study received considerable pedagogical support (e.g., pre-service training, lesson scripts, instructional coaches), so we cannot attribute their impact to their knowledge. Yet, if this intervention were to be taken to scale, fellows would likely receive such supports, so the parameter of policy interest is the estimated effect of combining fellows and supports.

Our results also add nuance to education-policy debates on the merits on lesson scripts. Opponents often argue that any gains from standardizing instruction must be weighted against the ensuing losses to teachers' professional autonomy (Valencia et al., 2006; Dresser, 2012). Yet, we found that the vast majority of fellows adhered to lesson scripts, suggesting that they may be a useful complement even for individuals with high levels of subject-matter expertise. In fact, the fellows who performed best in the assessment of content and pedagogical knowledge were less likely to omit, change, or add to the material in such scripts, implying that the fellows who were most likely to exercise discretion were precisely those least well equipped to do so. Future research should explore the extent to which this pattern is observed in other settings, and if so, how scripts can best encourage instructors to leverage their expertise appropriately.

Having demonstrated that this approach works, a logical next question is whether it could be implemented with comparable levels of fidelity and improve student achievement at scale. Of the reasons identified by List (2022) for why interventions that are successful in efficacy trials fail to sustain gains at scale, we are least concerned with the credibility of the evidence. Our randomized evaluation offers one of the most rigorous studies to date demonstrating that hiring individuals with subject-matter expertise can improve student achievement in a LMIC. The fact that we relied on a convenience sample, however, suggests that this intervention has more chances of scaling in urban, medium-sized, English-language public and charter schools. The motivation and capacity of the implementing organization seems much harder to replicate. SEI not only recruited, selected, and supported fellows; it also established partnership with the

[45] While fellows were expected to teach alongside regular teachers, as we show in section 5, the latter were typically not engaged in instructional tasks while the former were teaching.

local governments to place them in schools and with teacher-training colleges for certification; and it created detailed lesson scripts and hired instructional coaches for pedagogical support.[46]

Yet, perhaps the most important challenge to take this intervention to scale is to identify enough college students in math and science who are willing and able to become teachers. According to data for the 2020-2021 year, there were 32.7 million undergraduate students in India, 16% of whom studied science and 12% of whom were studying engineering (MoE, 2022). In fact, science and engineering were the second and fourth most popular majors, respectively. By comparison, there were 247,236 students in teacher-training colleges, which means that there were 21 science majors and 16 engineer majors per aspiring teacher across the country, suggesting that there is not a shortage of college students who can teach math and science.[47] Female enrollment in science and engineering is relatively high (52% and 29%, respectively), so getting more undergraduate students in these fields to enter teaching would not necessarily alter the sex breakdown in the profession. The question is whether they would *want* to teach. We do not have data on the career preferences of science and engineering majors across India. The only comparable program (Teach for India) has recruited 4,000 fellows since 2008,[48] but unlike the SEI fellowship, this program does not intend to keep its graduates in the profession. These figures, however, suggest that there is scope to increase the number of undergraduates in math and science who could enter teaching and raise the achievement of Indian students.

[46] Although this combination of know-how may be rare on a global scale, some of India's most successful education organizations already engage in similar tasks (e.g., Pratham, Educational Initiatives, Central Square Foundation, and their partner institutions), so such talent may be easier to find within the country.

[47] This is true even if we include students in education, who account for 5.3% of undegraduates nationwide.

[48] See Teach for India's website: `https://www.teachforindia.org/fellowship`.

References

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

Ajaja, O. P. and O. U. Eravwoke (2010). Effects of cooperative learning strategy on junior secondary school students achievement in integrated science. *The Electronic Journal for Research in Science & Mathematics Education 14*(1).

Akyeampong, T., T. Andrabi, A. 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).

Albornoz, F., M. V. Anuati, M. Furman, M. Luzuriaga, M. E. Podestá, and I. Taylor (2020). Training to teach science: Experimental evidence from Argentina. *World Bank Economic Review 34*(2), 393–417.

Alderman, H., J. R. Behrman, D. R. Ross, and R. H. Sabot (1996). The returns to endogenous human capital in Pakistan's rural wage labour market. *Oxford Bulletin of Economics and Statistics 58*(1), 29–55.

Alfonso, M., A. Santiago, and M. Bassi (2010). Estimating the impact of placing top university graduates in vulnerable schools in Chile. (Technical Note No. IDB-TN-230). Washington, DC: Inter-American Development Bank (IDB).

Altonji, J. G. and C. R. Pierret (2001). Employer learning and statistical discrimination. *Quarterly Journal of Economics 116*(1), 313–350.

Angrist, J. D. and V. Lavy (1997). The effect of a change in language of instruction on the returns to schooling in Morocco. *Journal of Labor Economics 15*(1), S48–S76.

Araujo, M. C., P. M. Carneiro, Y. Cruz-Aguayo, and N. Schady (2016). Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics 131*(3), 1415–1453.

Bando, R., E. Naslund-Hadley, and P. Gertler (2019). Effect of inquiry and problem based pedagogy on learning: Evidence from 10 field experiments in four countries. (NBER Working Paper No. 26280). Cambridge, MA: National Bureau of Economic Research (NBER).

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

Barro, R. J. (2001). Human capital and growth. *American Economic Review 91*(2), 12–17.

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.

Behrman, J. R., D. R. Ross, and R. H. Sabot (2008). Improving quality versus increasing the quantity of schooling: Estimates of rates of return from rural Pakistan. *Journal of Development Economics 85*(1-2), 94–104.

Berlinski, S. and M. Busso (2017). Challenges in educational reform: An experiment on active learning in mathematics. *Economic Letters 156*, 172–175.

Berry, J. and P. Mukherjee (2019). Pricing of private education in urban India: Demand, use, and impact. *Unpublished manuscript*. Athens, GA: University of Georgia.

Beuermann, D. W., E. Naslund-Hadley, I. J. Ruprah, and J. Thompson (2013). The pedagogy of science and environment: Experimental evidence from Peru. *The Journal of Development Studies 49*(5), 719–736.

Bhattacharjea, S., W. Wadhwa, and R. Banerji (2011). *Inside primary schools: A study of teaching and learning in rural India*. Mumbai, Maharashtra: Pratham.

Bianchi, N. and M. Giorcelli (2020). Scientific education and innovation: From technical diplomas to university stem degrees. *Journal of the European Economic Association 18*(5), 2608–2646.

Bishop, J. H. (1989). Is the test score decline responsible for the productivity growth decline? *American Economic Review 79*(1), 178–197.

Boissiere, M., J. B. Knight, and R. H. Sabot (1985). Earnings, schooling, ability, and cognitive skills. *American Economic Review 75*(5), 1016–1030.

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.

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

Chaudhury, N., J. Hammer, M. Kremer, K. Muralidharan, and F. H. Rogers (2006). Missing in action: Teacher and health worker absence in developing countries. *The Journal of Economic Perspectives 20*(1), 91–116.

Cruz-Aguayo, Y., P. Ibarrarán, and N. Schady (2017). Do tests applied to teachers predict their effectiveness? *Economics Letters 159*, 108–111.

de Barros, A. (2022). Explaining the productivity paradox: Experimental evidence from educational technology. *Unpublished manuscript*. Cambridge, MA: Massachusetts Institute of Technology (MIT).

Dresser, R. (2012). The impact of scripted literacy instruction on teachers and students. *Issues in Teacher Education 21*(1), 71–87.

Eble, A., C. Frost, A. Camara, B. Bouy, M. Bah, M. Sivaraman, P.-T. J. Hsieh, C. Jayanty, T. Brady, and P. Gawron (2021). How much can we remedy very low learning levels in rural parts of low-income countries? impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the gambia. *Journal of Development Economics 148*, 102539.

Gitomer, D. H., G. Phelps, B. H. Weren, H. Howell, and A. J. Croft (2014). *Evidence on the validity of content knowledge for teaching assessments*. San Francisco, CA: Jossey-Bass.

Glewwe, P. (1996). The relevance of standard estimates of rates of return to schooling for education policy: A critical assessment. *Journal of Development Economics 51*(2), 267–290.

Gray-Lobe, G., A. Keats, M. Kremer, I. Mbiti, and O. W. Ozier (2022). Can education be standardized? Evidence from Kenya. (Working Paper No. 2022-68). Chicago, IL: Becker Friedman Institute For Economics.

Grogger, J. and E. Eide (1995). Changes in college skills and the rise in the college wage premium. *Journal of Human Resources*, 280–310.

Hanushek, E. A. (2002). *Publicly provided education*, pp. 2045–2141. Amsterdam, the Netherlands; London, UK; and New York, NY: Elsevier Science, North Holland.

Hanushek, E. A. and D. D. Kimko (2000). Schooling, labor-force quality, and the growth of nations. *American Economic Review 90*(5), 1184–1208.

Hanushek, E. A., M. Piopiunik, and S. Wiederhold (2019). The value of smarter teachers: International evidence on teacher cognitive skills and student performance. *Journal of Human Resources 54*(4), 857–899.

Hanushek, E. A. and L. Woessmann (2008). The role of cognitive skils in economic development. *Journal of Economic Literature 46*(3), 607–668.

Hill, H. C., B. Rowan, and D. L. Ball (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal 42*(2), 371–406.

IEA (2015). PIRLS 2016: Assessment framework. TIMSS & PIRLS International Study Center. Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).

IEA (2017). TIMSS 2019: Assessment frameworks. Edited by Mullis, I. V. S. & Martin, M. O. TIMSS & PIRLS International Study Center. Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).

Jain, T., A. Mukhopadhyay, N. Prakash, and R. Rakesh (2022). Science education and labor market outcomes in a developing economy. *Economic Inquiry 60*(2), 741–763.

Jamison, E. A., D. T. Jamison, and E. A. Hanushek (2007). The effects of education quality on income growth and mortality decline. *Economics of Education Review 26*(6), 771–788.

Jolliffe, D. (1998). Skills, schooling, and household income in Ghana. *The World Bank Economic Review 12*(1), 81–104.

Kane, T. J., D. F. McCaffrey, T. Miller, and D. O. Staiger (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Measures of Effective Teaching Project*. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review 27*(6), 615–631.

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research (NBER).

Kane, T. J. and D. O. Staiger (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. *Measures of Effective Teaching Project*. Seattle, WA: Bill and Melinda Gates Foundation.

Kremer, M., N. Chaudhury, F. H. Rogers, K. Muralidharan, and J. Hammer (2005). Teacher absence in India: A snapshot. *Journal of the European Economic Association 3*(2-3), 658–667.

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review 10*(3), 179–214.

List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale*. New York, NY: Currency.

Loyalka, P., A. Popova, G. Li, and Z. Shi (2019). Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics 11*(3), 128–154.

Lucas, A. M., P. J. McEwan, M. Ngware, and M. Oketch (2014). Improving early-grade literacy in east africa: Experimental evidence from kenya and uganda. *Journal of Policy Analysis and Management 33*(4), 950–976.

Mbiti, I. M., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani (2019). Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. *The Quarterly Journal of Economics 134*(3), 1627–1673.

Metzler, J. and L. Woessmann (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics 99*(2), 486–496.

MoE (2022). All India survey on higher education 2020-2021. New Delhi, India: Department of Higher Education, Minsitry of Education, Government of India.

Molina, E., S. F. Fatima, A. D. Ho, C. Melo, T. Wilichowski, and A. Pushparatnam (2020). Measuring the quality of teaching practices in primary schools: Assessing the validity of the Teach observation tool in Punjab, Pakistan. *Teaching and Teacher Education 96*, 103171.

Mulligan, C. B. (1999). Galton versus the human capital approach to inheritance. *Journal of Political Economy 107*(SG), S184–S224.

Muralidharan, K., J. Das, A. Holla, and A. Mohpal (2017). The fiscal cost of weak governance: Evidence from teacher absence in India. *Journal of Public Economics 145*(C), 116–135.

Murnane, R. J., J. B. Willett, M. J. Braatz, and Y. Duhaldeborde (2001). Do different dimensions of male high school students' skills predict labor market success a decade later? *Economics of Education Review 20*(4), 311–320.

Murnane, R. J., J. B. Willett, and F. Levy (1995). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics*, 251–266.

Naslund-Hadley, E., S. W. Parker, and J. M. Hernandez-Agramonte (2014). Fostering early math comprehension: Experimental evidence from Paraguay. *Global Education Review 1*, 135–154.

NCERT (2018). National achievement survey 2017: District report cards. New Delhi, India: National Council of Educational Research and Training (NCERT).

NCES (2018). Digest of education statistics 2018. U.S. Department of Education, National Center for Education Statistics, Common Core of Data. URL: https://nces.ed.gov/programs/digest/d18/tables/dt18_215.30.asp?current=yes. Last accessed: March 7, 2020.

Neal, D. A. and W. R. Johnson (1996). The role of premarket factors in Black-White wage differences. *Journal of Political Economy 104*(5), 869–895.

NIEPA (2017). Elementary education in India: Where do we stand? (District report cards 2016-2017, Vol. I). New Delhi, India: National Institute of Educational Planning and Administration (NIEPA).

OECD (2019). TALIS 2018 results (Vol. I): Teachers and school leaders as lifelong learners. Paris, France: Organisation for Economic Co-operation and Deveopment (OECD).

OECD (2022). Education at a Glance 2022: OECD Indicators. Paris, France: Organisation for Economic Co-operation and Deveopment.

O'Neill, J. (1990). The role of human capital in earnings differences between black and white men. *Journal of Economic Perspectives 4*(4), 25–45.

Piper, B. and M. Korda (2010). EGRA Plus: Liberia. Program evaluation report. Prepared for USAID/Liberia under the Education Data for Decision Making (EdData II) project, Early Grade Reading Assessment (EGRA): Plus Project, Task Order No. EHC-E-06-04-00004-00 (RTI Task 6). Research Triangle Park, NC: RTI International.

Piper, B., S. S. Zuilkowski, M. Dubeck, E. Jepkemei, and S. J. King (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development 106*, 324–336.

Popova, A., D. K. Evans, M. E. Breeding, and V. Arancibia (2022). Teacher professional development around the world: The gap between evidence and practice. *The World Bank Research Observer 37*(1), 107–136.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 247–252.

Rockoff, J. E., B. A. Jacob, T. J. Kane, and D. O. Staiger (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy 6*(1), 43–74.

Romero, M., J. Sandefur, and W. A. Sandholtz (2020). Outsourcing education: Experimental evidence from Liberia. *American Economic Review 110*(2), 364–400.

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.

Santibanez, L. (2006). Why we should care if teachers get A's: Teacher test scores and student achievement in Mexico. *Economics of Education Review 25*(5), 510–520.

Stallings, J. A. (1977). *Learning to look: A handbook on classroom observation and teaching models*. Wadsworth Pub. Co., Belmont.

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.

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: International Evaluation Association (IEA).

Tilson, T., A. Kamlongera, M. Pucilowski, and D. Nampota (2013). Evaluation of the Malawi Professional Development Support (MTPDS) Program: Final evaluation report. Prepared for USAID under Requisition No. REQ-612-12-000029. Washington, DC: Social Impact, Inc.

UN (2019). World population prospects, 2019 revision. New York, NY: United Nations Department of Economic and Social Affairs, Population Dynamics. URL: https://population.un.org/wpp/Publications/. Last accessed: March 7, 2020.

UNESCO (2022). Global education monitoring report 2021/2. Non-state actors in education. Who chooses? Who loses? Paris, France: United Nations Educational, Scientific, and Cultural Organization (UNESCO).

Valencia, S. W., N. A. Place, S. D. Martin, and P. L. Grossman (2006). Curriculum materials for elementary reading: Shackles and scaffolds for four beginning teachers. *The Elementary School Journal 107*(1), 93–120.

von Hippel, P. T. and L. Bellows (2018). How much does teacher quality vary across teacher preparation programs? reanalyses from six states. *Economics of Education Review 64*, 298–312.

Wachanga, S. W. and J. G. Mwangi (2004). Effects of the cooperative class experiment teaching method on secondary school students' chemistry achievement in Kenya's Nakuru district. *International Education Journal 5*(1), 26–36.

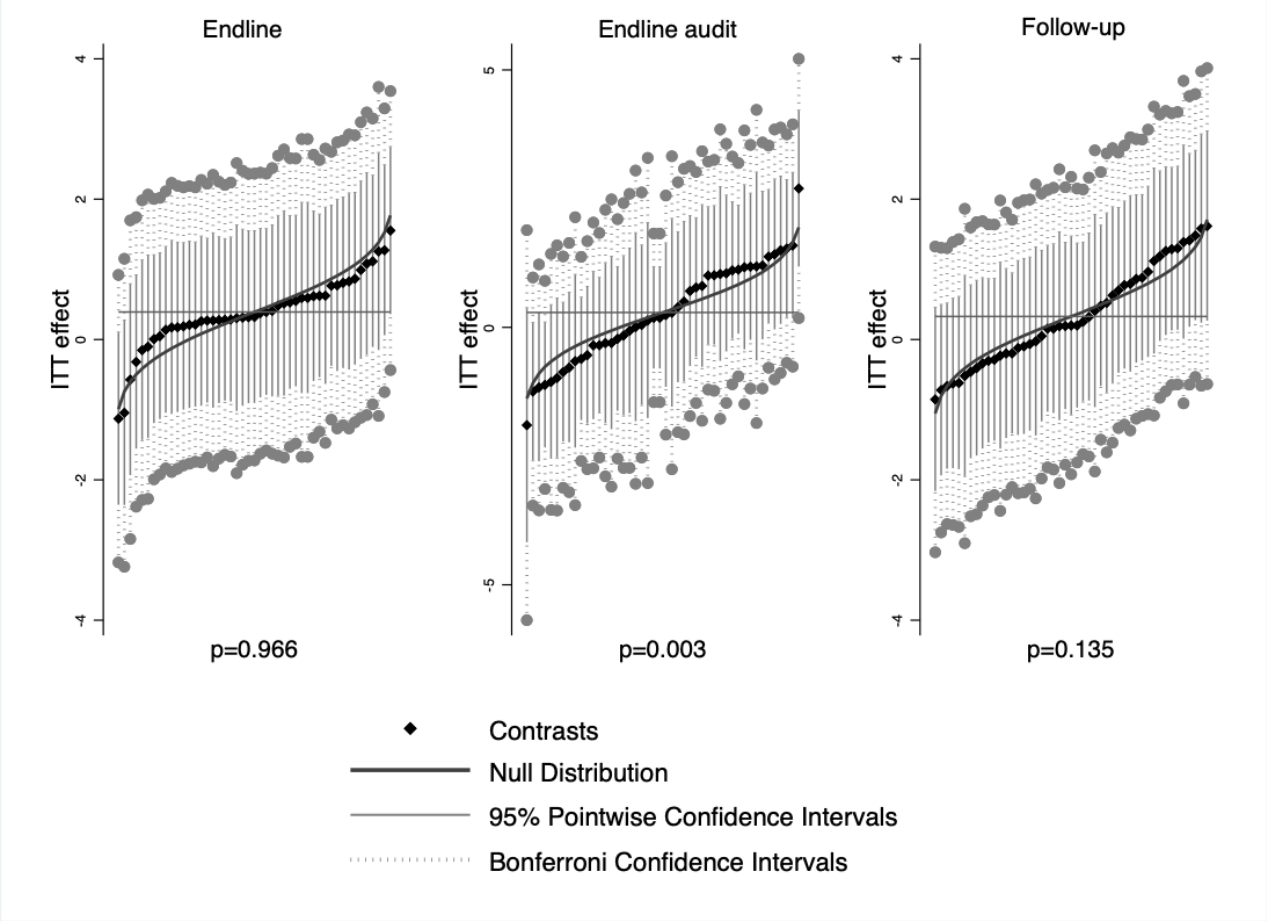
Woessmann, L. (2003). Specifying human capital. *Journal of Economic Surveys 17*(3), 239–270.

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 (2023). Education statistics (Edstats). `https://datatopics.worldbank.org/education/` Retrieved: June 8, 2023.

Yen, W. M. and A. R. Fitzpatrick (2006). Item response theory. In Brennan, R. (Ed.) *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Figure 1: Heterogeneous impact on standardized test scores by school (endline, endline audit, and follow-up)



Notes: The figure provides a "caterpillar plot" (von Hippel and Bellows, 2018) of impacts on the first principal component of a principal-component analysis of standardized scores in math and science by school at endline, about nine months after the rollout of the intervention (April 2018); the endline "audit," about nine months after the rollout of the intervention (April 2018); and follow-up, about 11 months after the rollout of the intervention (June 2018). Each black dot refers to the point estimate for a given school. Bonferroni confidence intervals adjust standard errors for multiple hypothesis testing. The black solid line shows the null distribution of "effects" that can be expected due to error. The p-values are for a test of the null hypothesis of no heterogeneity.

Table 1: Differences between student characteristics (baseline)

	(1) Control	(2) Treatment	(3) Col. (2)-(1)	(4) N
Female	0.506	0.542	0.036	2,581
	[0.500]	[0.498]	(0.031)	
Age	10.781	10.852	0.055	2,404
	[1.011]	[1.038]	(0.048)	
Speaks Marathi at home	0.690	0.690	-0.000	2,581
	[0.463]	[0.463]	(0.023)	
Speaks Hindi at home	0.169	0.164	-0.005	2,581
	[0.375]	[0.370]	(0.018)	
Speaks English at home	0.011	0.006	-0.004	2,581
	[0.102]	[0.080]	(0.005)	
Mother completed primary school	0.634	0.600	-0.034	2,657
	[0.482]	[0.490]	(0.023)	
Father completed primary school	0.768	0.778	0.009	2,657
	[0.422]	[0.416]	(0.019)	
Student has a desk to study	0.279	0.257	-0.022	2,581
	[0.449]	[0.437]	(0.029)	
Student has own room	0.172	0.204	0.032	2,581
	[0.378]	[0.403]	(0.030)	
Student has a computer	0.195	0.176	-0.019	2,579
	[0.396]	[0.381]	(0.025)	
Student has Internet	0.425	0.373	-0.053	2,579
	[0.495]	[0.484]	(0.045)	
Student has a TV	0.901	0.894	-0.006	2,581
	[0.299]	[0.307]	(0.017)	
Attends tuition in math	0.347	0.315	-0.032	2,288
	[0.476]	[0.465]	(0.036)	
Attends tuition in science	0.260	0.228	-0.032	2,267
	[0.439]	[0.420]	(0.042)	
Attends tuition in Marathi or Hindi	0.368	0.356	-0.012	2,293
	[0.482]	[0.479]	(0.027)	
Attends tuition in English	0.327	0.305	-0.022	2,353
	[0.469]	[0.461]	(0.034)	
Math (standardized score)	0.000	-0.131	-0.131*	2,657
	[1.000]	[0.955]	(0.068)	
Science (standardized score)	-0.000	-0.050	-0.050	2,657
	[1.000]	[1.011]	(0.066)	
Language (standardized score)	0.000	-0.066	-0.065	2,657
	[1.000]	[1.016]	(0.060)	
Raven's matrices (standardized score)	0.000	0.045	0.045	2,657
	[1.000]	[0.984]	(0.055)	

Notes: This table compares the students in the control and treatment groups at baseline. It shows the means and standard deviations for each group (columns 1-2) and tests for differences between groups including randomization-strata (i.e., grade) fixed effects (column 3). The sample includes all students observed at baseline. Standard deviations appear in brackets, and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2: Differences between teacher characteristics (endline)

	(1)	(2)	(3)	(4)	(5)
	PMC teachers		Diff. between PMC teachers	SEI fellows	Diff. between PMC teachers and SEI fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
Female	0.809	0.711	-0.101	0.633	-0.078
	[0.398]	[0.458]	(0.085)	[0.487]	(0.107)
Age	41.553	41.867	0.250	20.816	-20.989***
	[8.304]	[8.033]	(1.553)	[1.811]	(1.231)
Years of teaching experience (total)	18.936	17.578	-1.472	1.184	-16.307***
	[8.578]	[8.286]	(1.601)	[0.486]	(1.231)
Years of teaching experience (at this school)	5.191	5.089	-0.130	1.000	-4.085***
	[4.808]	[4.039]	(0.993)	[0.000]	(0.638)
Years of math teaching experience (at this school)	1.511	2.533	1.007**	1.000	-1.524***
	[1.266]	[2.693]	(0.467)	[0.000]	(0.395)
Years of science teaching experience (at this school)	1.489	2.533	1.029**	1.000	-1.524***
	[1.249]	[2.693]	(0.467)	[0.000]	(0.395)
Has bachelor's degree or higher	0.809	0.867	0.056	0.347	-0.520***
	[0.398]	[0.344]	(0.076)	[0.481]	(0.080)
Also teaches English	0.957	1.000	0.043	0.000	-1.000
	[0.204]	[0.000]	(0.030)	[0.000]	(0.000)
Also teaches Indian languages	0.894	1.000	0.107**	0.000	-1.000
	[0.312]	[0.000]	(0.045)	[0.000]	(0.000)
Teaches in English	0.064	0.022	-0.043	0.041	0.017
	[0.247]	[0.149]	(0.030)	[0.200]	(0.020)
Teaches in Hindi	0.043	0.178	0.136**	0.163	-0.015
	[0.204]	[0.387]	(0.052)	[0.373]	(0.012)
Teaches in Marathi	0.894	0.800	-0.093*	0.796	-0.002
	[0.312]	[0.405]	(0.051)	[0.407]	(0.021)
Teacher assessment (proportion-correct score)	0.643	0.621	-0.022	0.771	0.150***
	[0.099]	[0.096]	(0.019)	[0.077]	(0.017)
Instructional practices (proportion-correct score)	0.308	0.280	-0.028	0.493	0.213***
	[0.165]	[0.179]	(0.032)	[0.163]	(0.038)
Student errors (proportion-correct score)	0.543	0.503	-0.041	0.689	0.187***
	[0.166]	[0.154]	(0.030)	[0.165]	(0.030)
Content knowledge (proportion-correct score)	0.836	0.829	-0.007	0.931	0.102***
	[0.120]	[0.122]	(0.027)	[0.071]	(0.022)
N (instructors)	47	45	92	49	94

Notes: This table compares the Pune Municipal Corporation (PMC) teachers and Science Education Initiative (SEI) fellows at endline. It shows the means and standard deviations for each group and and tests for differences between groups including randomization-strata (i.e., grade) fixed effects). The sample includes all teachers observed at endline. Standard deviations appear in brackets, and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: Timeline of the study

(1)	(2)	(3)	(4)	(5)	(6)
			Student participation rates		
Month	Event	Grades	Total classrooms	Control classrooms	Treatment classrooms
A. 2017-2018 school year					
June	School year starts				
July	Student assessments	5, 6	100%	100%	100%
	Student surveys	5, 6	100%	100%	100%
December	Announced classroom observations	5, 6	99%	100%	98%
February	Unannounced school visits	5, 6	99%	100%	98%
	Administrative data on student attendance	5, 6	100%	100%	100%
March-	Student assessments	5, 6			
April	– Math and science		88%	87%	89%
	– Language		86%	85%	87%
	Student surveys	5, 6	80%	79%	81%
	Student "audit" assessments of math and science	5, 6	21%	22%	22%
April-	Teacher surveys	5, 6	95%	93%	98%
May	Teacher assessments	5, 6	94%	92%	98%
B. 2018-2019 school year					
June	School year starts				
June-	Student assessments	6, 7			
August	– Math		81%	81%	81%
	– Science		80%	80%	81%
	– Language		81%	80%	81%
	Administrative data on student attendance	6, 7			

Notes: The table shows the timeline for the interventions and rounds of data collection for the study, including the month in which each event occurred (column 1), a brief description of the event (column 2), the target grades (column 3), and the percentage of students that participated in each event by experimental group (columns 4-6). The student "audit" assessments only targeted 25% of the study sample.

Table 4: Impact on standardized test scores (endline, endline audit, and follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.247***	0.340***	0.207***	0.216***	0.132**	0.151***
	(0.048)	(0.048)	(0.051)	(0.049)	(0.052)	(0.039)
Baseline score		0.646***		0.485***		0.555***
		(0.022)		(0.030)		(0.026)
N (students)	2524	2307	2524	2307	2524	2307
B. Endline audit						
Treatment	0.046	0.091	0.367***	0.399***		
	(0.106)	(0.104)	(0.116)	(0.113)		
Baseline score		0.139**		0.203***		
		(0.061)		(0.061)		
N (students)	553	551	553	551		
C. Follow-up						
Treatment	0.283***	0.356***	0.120**	0.142**	0.059	0.081**
	(0.051)	(0.046)	(0.046)	(0.053)	(0.042)	(0.037)
Baseline score		0.651***		0.321***		0.446***
		(0.028)		(0.028)		(0.029)
N (students)	2369	2023	2369	2023	2369	2023

Notes: This table compares students' standardized scores in the control and treatment groups based on assessments administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018) and at follow-up, about 11 months after the rollout of the intervention (June 2018). Scores are standardized to have mean zero and standard deviation one in the control group. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: Impact on proportion-correct repeated and non-repeated items (endline)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. All items						
Treatment	0.049***	0.067***	0.036***	0.038***	0.031**	0.035***
	(0.010)	(0.009)	(0.009)	(0.009)	(0.012)	(0.009)
Baseline score		0.584***		0.509***		0.520***
		(0.019)		(0.031)		(0.024)
Control mean						
N (students)	2524	2307	2524	2307	2524	2307
B. Repeated items						
Treatment	0.071***	0.094***	0.054***	0.058***	0.031**	0.036***
	(0.012)	(0.012)	(0.011)	(0.011)	(0.013)	(0.010)
Baseline score		0.749***		0.619***		0.522***
		(0.024)		(0.036)		(0.027)
Control mean	0.542		0.462		0.539	
N (students)	2524	2307	2524	2307	2524	2307
C. Non-repeated items						
Treatment	0.028***	0.041***	0.018**	0.017**	0.030**	0.034***
	(0.008)	(0.008)	(0.009)	(0.008)	(0.013)	(0.010)
Baseline score		0.428***		0.393***		0.517***
		(0.018)		(0.030)		(0.024)
Control mean	0.382		0.388		0.512	
N (students)	2524	2307	2524	2307	2524	2307
P-value of the difference	0.000	0.000	0.933	0.000	0.000	0.862

Notes: This table compares students' proportion-correct scores on items in both the baseline and endline assessments ("repeated") and on items only in the endline assessments ("non-repeated") across the control and treatment groups, about nine months after the rollout of the intervention (April 2018). Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. As Table 4 indicates, the mean standardized effects on all items without accounting for baseline performance are 0.247 SDs ($p < 0.01$) in math, 0.207 SDs ($p < 0.01$) in science, and 0.132 ($p < 0.05$) in language. The corresponding effects for repeated items are 0.275 SDs ($p < 0.01$) in math, 0.251 SDs ($p < 0.01$) in science, and 0.125 SDs ($p < 0.05$) in language. The corresponding effects for non-repeated items are 0.169 SDs ($p < 0.01$) in math, 0.104 SDs ($p < 0.05$) in science, and 0.124 SDs ($p < 0.05$) in language. Standardizing with respect to the total percent-correct score, effects for repeated items would be 0.359 SDs ($p < 0.01$) in math, 0.306 SDs ($p < 0.01$) in science, and 0.134 SDs ($p < 0.05$) in language; and those for non-repeated items are 0.142 SDs ($p < 0.01$) in math, 0.101 SDs ($p < 0.05$) in science, and 0.130 SDs ($p < 0.05$) in language.

Table 6: Impact on proportion-correct repeated and non-repeated items (follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. All items						
Treatment	0.058***	0.073***	0.019**	0.022**	0.006	0.008**
	(0.010)	(0.009)	(0.007)	(0.008)	(0.004)	(0.004)
Baseline score		0.609***		0.303***		0.186***
		(0.026)		(0.026)		(0.012)
Control mean						
N (students)	2369	2023	2369	2023	2369	2023
B. Repeated items						
Treatment	0.070***	0.087***	0.027**	0.032**	0.018	0.023**
	(0.013)	(0.012)	(0.011)	(0.013)	(0.012)	(0.010)
Baseline score		0.738***		0.439***		0.461***
		(0.033)		(0.038)		(0.030)
Control mean	0.531		0.380		0.476	
N (students)	2369	2023	2369	2023	2369	2023
C. Non-repeated items						
Treatment	0.041***	0.053***	0.015**	0.017**	0.005	0.011
	(0.010)	(0.009)	(0.006)	(0.006)	(0.009)	(0.011)
Baseline score		0.439***		0.200***		0.302***
		(0.023)		(0.023)		(0.024)
Control mean	0.350		0.280		0.413	
N (students)	2369	2023	2369	2023	2369	2023
P-value of the difference	0.012	0.315	0.292	0.027	0.259	0.426

Notes: This table compares students' proportion-correct scores on items in both the baseline and endline assessments ("repeated") and on items only in the endline assessments ("non-repeated") across the control and treatment groups, about nine months after the rollout of the intervention (April 2018). Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. As Table 4 indicates, the mean standardized effects on all items without accounting for baseline performance are 0.283 SDs ($p < 0.01$) in math, 0.120 SDs ($p < 0.05$) in science, and 0.059 ($p > 0.10$) in language. The corresponding effects for repeated items are 0.274 SDs ($p < 0.01$) in math, 0.110 SDs ($p < 0.05$) in science, and 0.073 SDs ($p > 0.10$) in language. The corresponding effects for non-repeated items are 0.227 SDs ($p < 0.01$) in math, 0.112 SDs ($p < 0.05$) in science, and 0.022 SDs ($p > 0.10$) in language. Standardizing with respect to the total percent-correct score, effects for repeated items would be 0.345 SDs ($p < 0.01$) in math, 0.169 SDs ($p < 0.05$) in science, and 0.177 SDs ($p > 0.10$) in language; and those for non-repeated items are 0.201 SDs ($p < 0.01$) in math, 0.097 SDs ($p < 0.05$) in science, and 0.051 SDs ($p > 0.10$) in language.

Table 7: Impact on instructional practices (announced observations)

	(1) PMC teachers	(2) SEI fellows	(3) Difference between PMC teachers and SEI fellows
	Control	Treatment	Col. (2)-(1)
A. Negative practices			
Share of instructors who...			
...stayed in one spot during the lesson	0.188	0.060	-0.127**
	[0.392]	[0.240]	(0.060)
...sat down for a long period of time	0.042	0.020	-0.022
	[0.201]	[0.141]	(0.029)
...used phone during lesson	0.052	0.080	0.028
	[0.223]	[0.274]	(0.049)
...were upset at incorrect answers	0.062	0.180	0.117*
	[0.243]	[0.388]	(0.062)
...hit, pinched, or slapped students	0.021	0.000	-0.021
	[0.144]	[0.000]	(0.014)
...shouted or used harsh language	0.031	0.020	-0.011
	[0.175]	[0.141]	(0.027)
Composite index (std.)	-0.000	0.174	0.174
	[1.000]	[0.919]	(0.172)
N (instructors)	96	50	146
B. Positive practices			
Share of instructors who...			
...made eye contact with nearly all students	0.750	0.740	-0.010
	[0.435]	[0.443]	(0.062)
...called on nearly all students by name	0.635	0.640	0.005
	[0.484]	[0.485]	(0.084)
...called on students from all seating rows	0.688	0.780	0.093
	[0.466]	[0.418]	(0.098)
...asked both closed and open questions	0.625	0.780	0.155*
	[0.487]	[0.418]	(0.088)
...asked students to explain their answers	0.469	0.760	0.291***
	[0.502]	[0.431]	(0.095)
...corrected student answers	0.719	0.940	0.221***
	[0.452]	[0.240]	(0.070)
...allowed students to ask questions	0.385	0.580	0.195**
	[0.489]	[0.499]	(0.095)
...assigned classwork	0.635	0.740	0.105
	[0.484]	[0.443]	(0.069)
...helped individual students	0.677	0.960	0.283***
	[0.470]	[0.198]	(0.064)
...summarized lesson at the end	0.271	0.360	0.089
	[0.447]	[0.485]	(0.091)
...assigned homework	0.479	0.640	0.161*
	[0.502]	[0.485]	(0.092)
...praise or encouraged students	0.760	0.960	0.200***
	[0.429]	[0.198]	(0.058)
...smiled, joked, or laughed	0.167	0.280	0.113
	[0.375]	[0.454]	(0.090)
Composite index (std.)	-0.000	0.730	0.730***
	[1.000]	[0.687]	(0.175)
N (instructors)	96	50	146

Notes: This table compares the instructional practices of Pune Municipal Corporation (PMC) teachers in the control and treatment groups and of Science Education Initiative (SEI) fellows in the treatment group, based on announced observations about five months after the rollout of the intervention (December 2017). Observations of PMC teachers lasted 30 minutes and those of SEI fellows lasted 120 minutes. Panel A displays results expressed as a proportion of class time and Panel B shows results in terms of minutes per lesson. The composite indexes are the first principal components of the variables in each panel, standardized with respect to the control group. Standard deviations appear in brackets and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 8: Impact on attendance and punctuality (unannounced school visits)

	(1)	(2)	(3)	(4)	(5)
	PMC teachers		Impact on PMC teachers	SEI fellows	Difference between PMC teachers and SEI fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Attendance and punctuality					
Share of instructors who...					
...attended today	0.840	0.878	0.038	0.720	-0.120
	[0.370]	[0.331]	(0.078)	[0.454]	(0.089)
...arrived on time today	0.740	0.714	-0.026	0.440	-0.300***
	[0.443]	[0.456]	(0.087)	[0.501]	(0.084)
N (instructors)	50	49	99	50	100
B. Location at the school					
Share of present instructors who...					
...were in the classroom	0.048	0.047	-0.001	0.556	0.508***
	[0.216]	[0.213]	(0.031)	[0.504]	(0.094)
...were in the principal's office	0.690	0.744	0.049	0.167	-0.523***
	[0.468]	[0.441]	(0.079)	[0.378]	(0.095)
...were elsewhere in the school	0.262	0.209	-0.048	0.278	0.015
	[0.445]	[0.412]	(0.074)	[0.454]	(0.098)
N (present instructors)	42	43	85	36	78

Notes: This table compares average attendance and punctuality of Pune Municipal Corporation (PMC) teachers in the control and treatment groups and of Science Education Initiative (SEI) fellows in the treatment group, based on unannounced visits about seven months after the rollout of the intervention (February 2018). Panel A displays results for all instructors and Panel B shows results only for instructors who were present on the day of the visit. Standard deviations appear in brackets and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

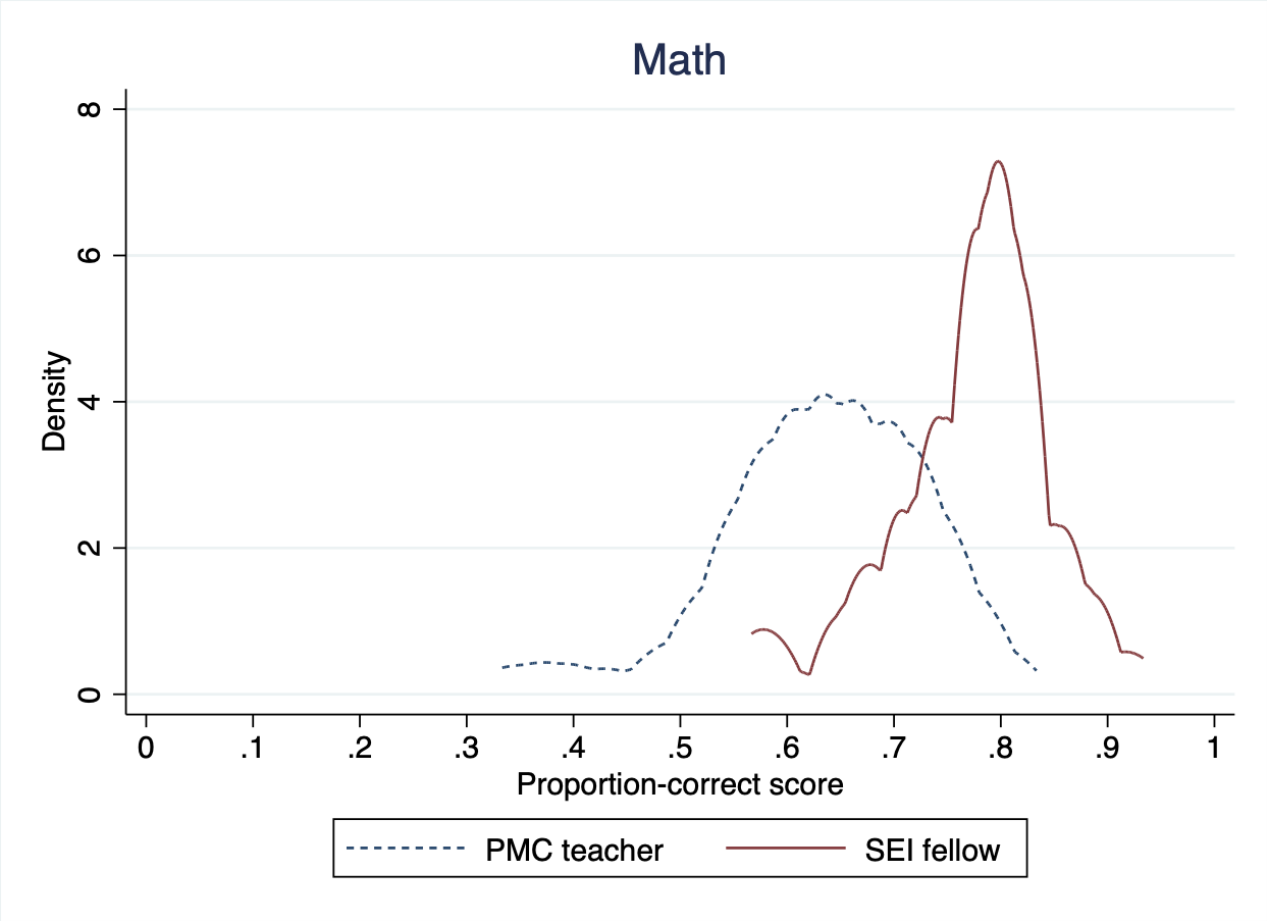
Table 9: Impact on allocation of instructional time (announced observations)

	(1)	(2)	(3)	(4)	(5)
	PMC teachers		Impact on PMC teachers	SEI fellows	Difference between PMC teachers and SEI fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time...					
...on task	0.778	0.080	-0.698***	0.745	-0.034
	[0.199]	[0.134]	(0.026)	[0.156]	(0.031)
...on class management	0.139	0.208	0.069*	0.159	0.021
	[0.134]	[0.264]	(0.041)	[0.126]	(0.021)
...off task	0.083	0.692	0.609***	0.096	0.013
	[0.135]	[0.342]	(0.048)	[0.112]	(0.021)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Minutes of lesson time...					
...on task	23.344	9.600	-13.744***	89.388	65.992***
	[5.956]	[16.081]	(2.295)	[18.665]	(2.720)
...on class management	4.156	24.960	20.804***	19.102	14.967***
	[4.022]	[31.688]	(4.293)	[15.083]	(2.161)
...off task	2.500	83.040	80.540***	11.510	9.041***
	[4.052]	[41.060]	(5.557)	[13.407]	(1.904)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of instructional time of Pune Municipal Corporation (PMC) teachers in the control and treatment groups and of Science Education Initiative (SEI) fellows in the treatment group, based on announced observations about five months after the rollout of the intervention (December 2017). Observations of PMC teachers lasted 30 minutes and those of SEI fellows lasted 120 minutes. Panel A displays results expressed as a proportion of lesson time and Panel B shows results in terms of minutes per lesson. Standard deviations appear in brackets and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

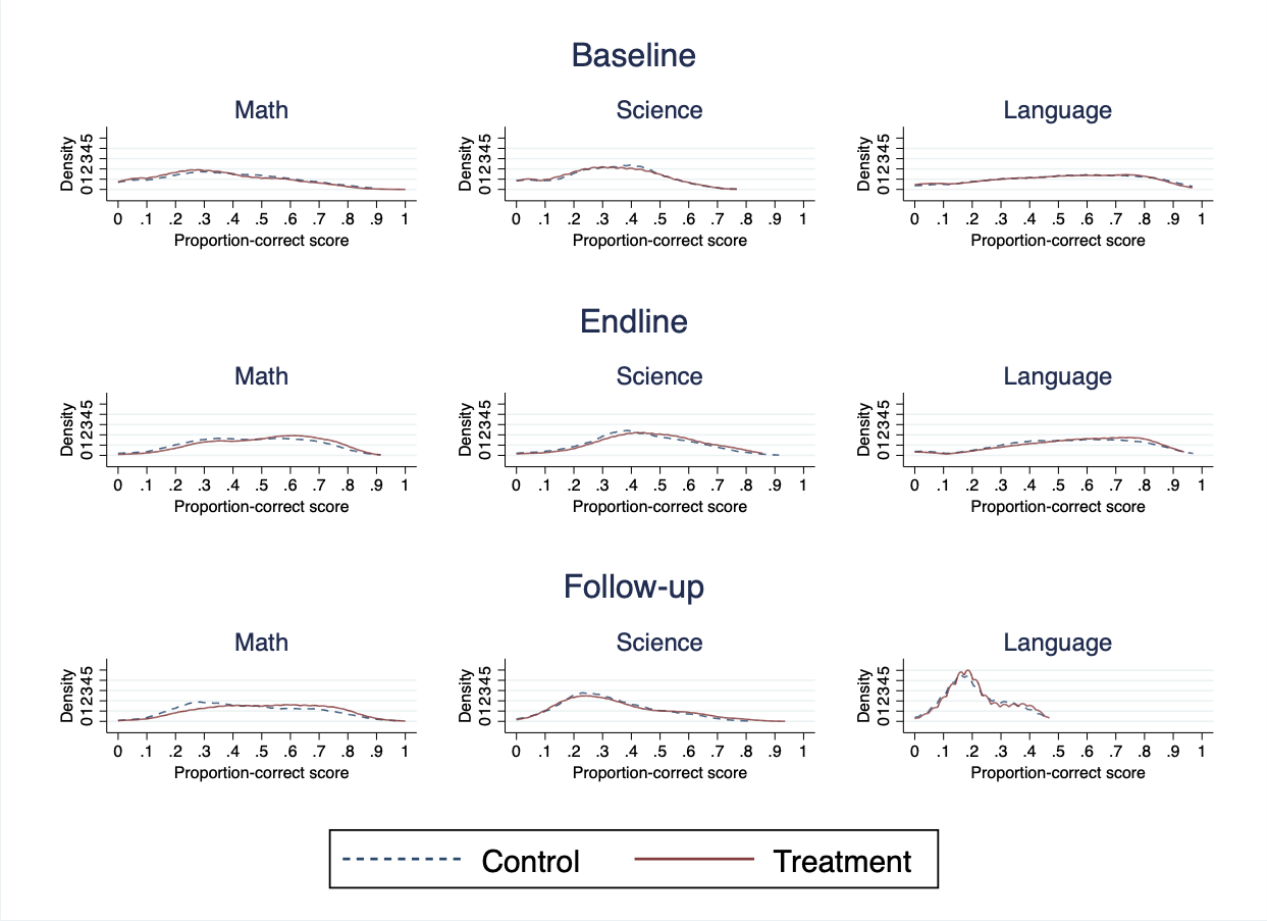
Appendix A Additional figures and tables

Figure A.1: Distributions of proportion-correct scores on teacher assessments



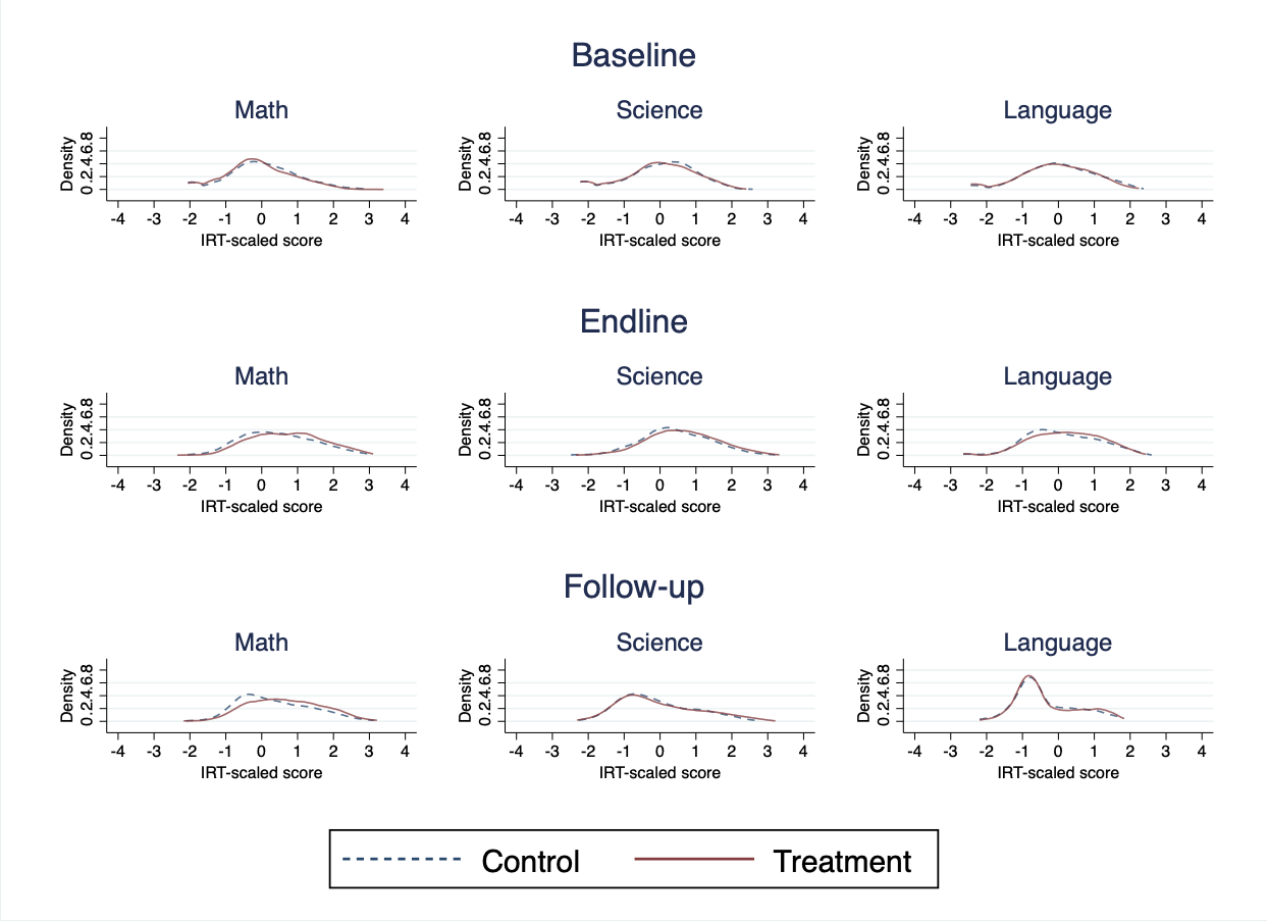
Notes: The figure shows the distribution of proportion-scaled scores on the teacher assessments. Proportion-correct scores indicate the proportion of items on each test answered correctly.

Figure A.2: Distributions of proportion-correct scores on student assessments



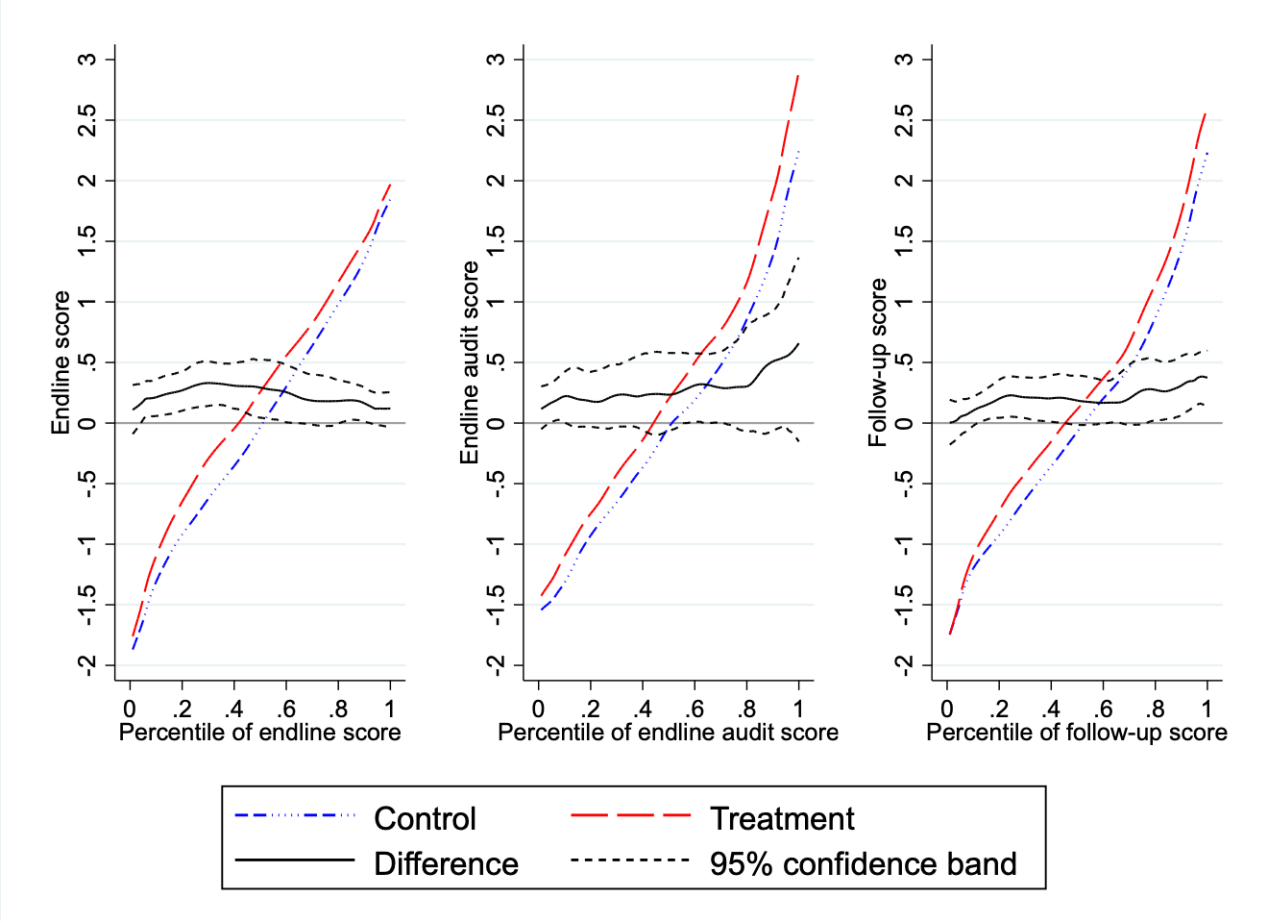
Notes: The figure shows the distribution of proportion-scaled scores on the student assessments. Proportion-correct scores indicate the proportion of items on each test answered correctly.

Figure A.3: Distributions of IRT-scaled scores on student assessments



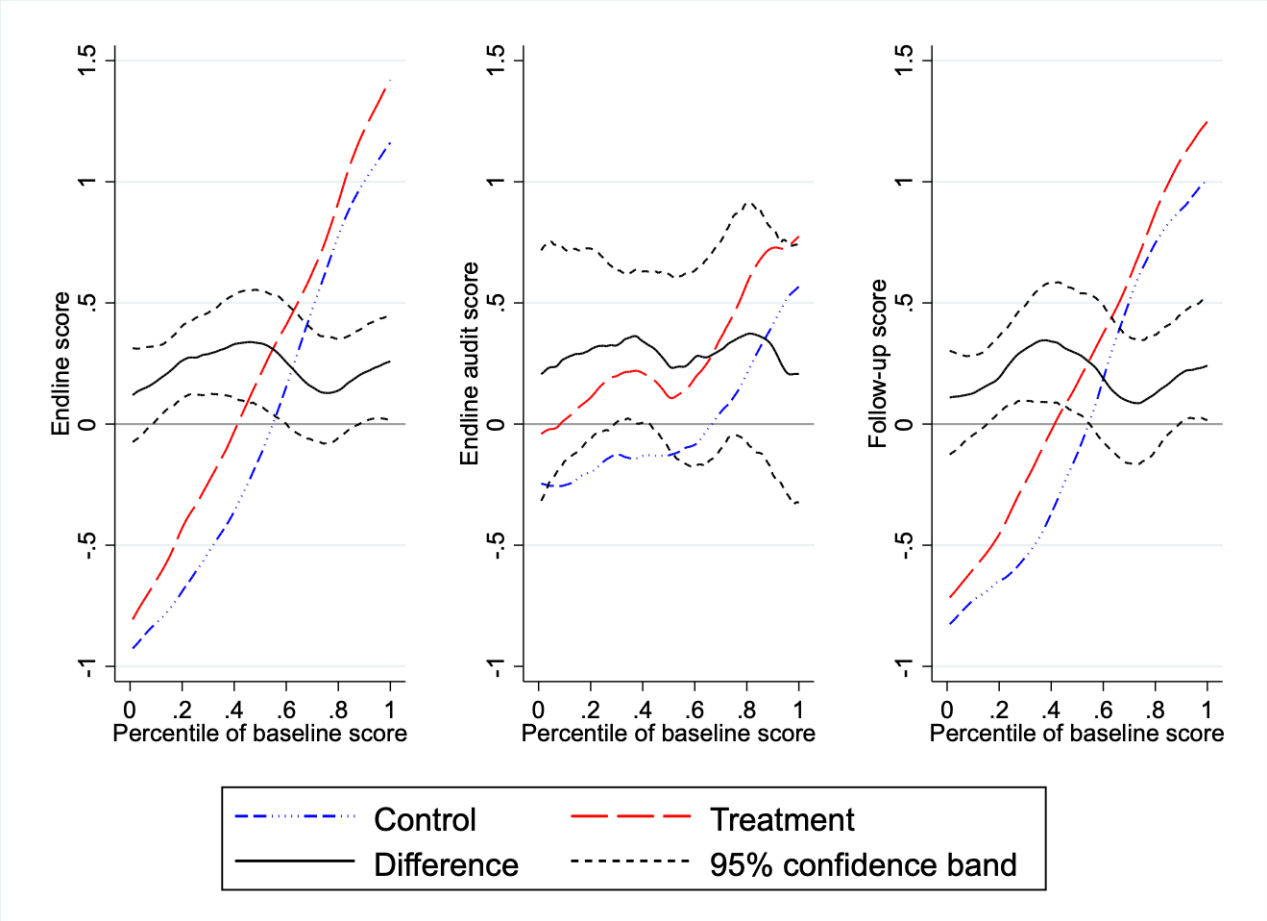
Notes: The figure shows the distribution of IRT-scaled scores on the student assessments. IRT-scaled scores are expressed in standard deviations.

Figure A.4: Quantile treatment effects (endline, endline audit, and follow-up)



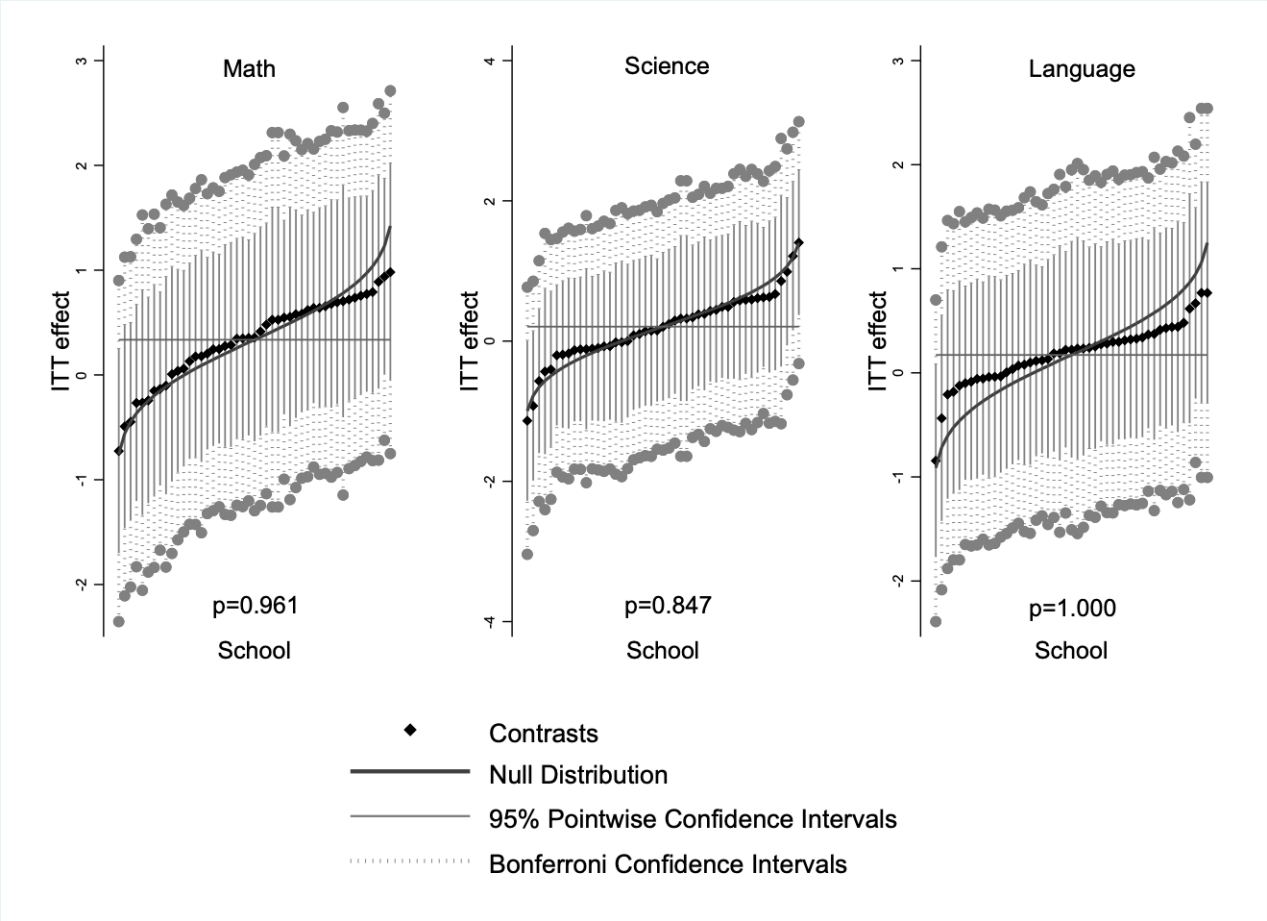
Notes: The figure shows quantiles of endline, endline audit, and follow-up composite scores (the first principal component of a principal component analysis of all subjects assessed at each round) for treatment and control students who participated in the baseline and each of these rounds, estimated by polynomial regressions of each round's scores on that round's percentiles separately by experimental group. Dashed black lines display bootstrapped 95% confidence intervals.

Figure A.5: Average treatment effects by baseline score (endline, endline audit, and follow-up)



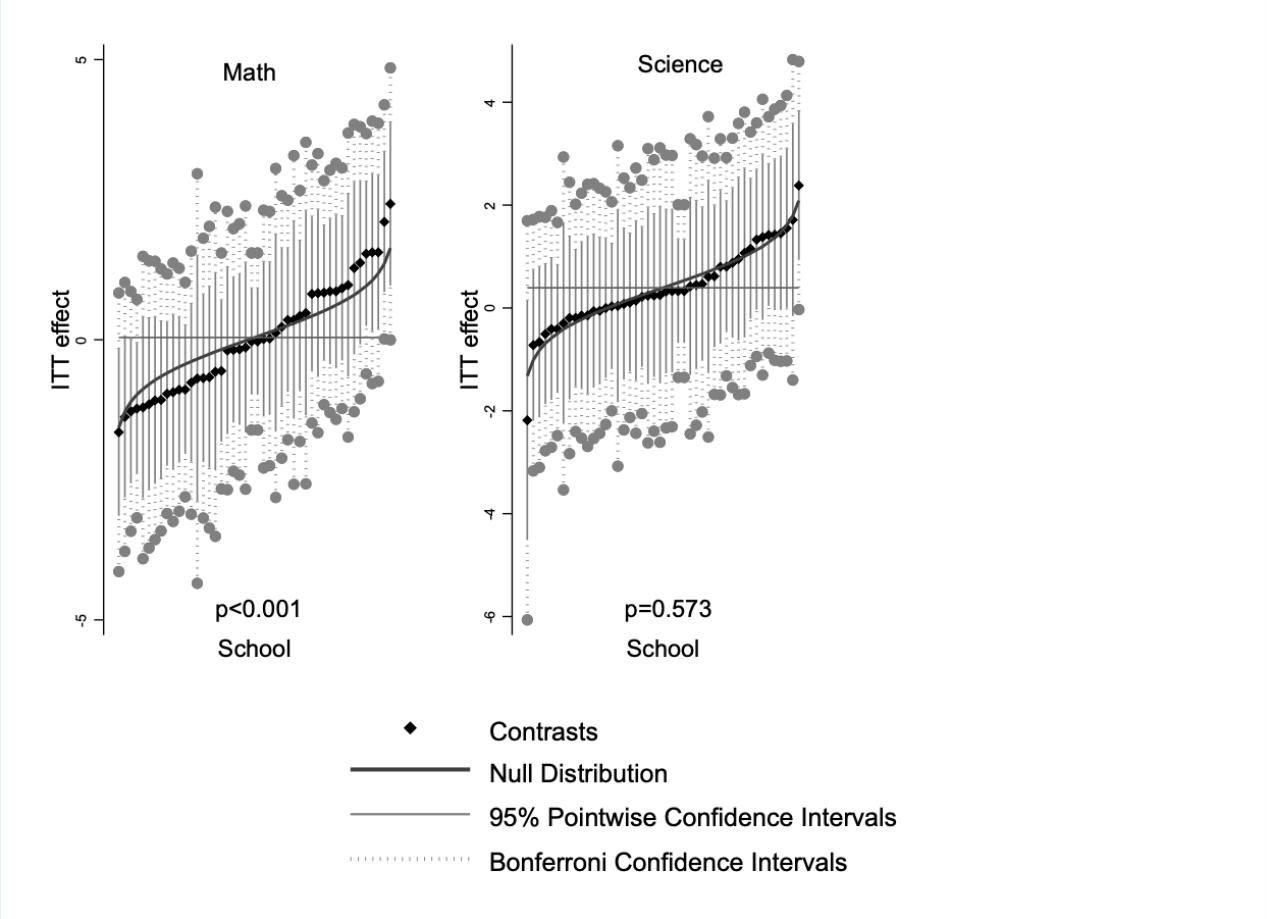
Notes: The figure shows estimates of average endline, endline audit, and follow-up composite scores (the first principal component of a principal component analysis of all subjects assessed at each round) and treatment effects at each percentile of baseline composite score for treatment and control students who participated in the baseline and each of these rounds, estimated by polynomial regression. Dashed black lines display bootstrapped 95% confidence intervals.

Figure A.6: Heterogeneous impact on standardized test scores by school (endline)



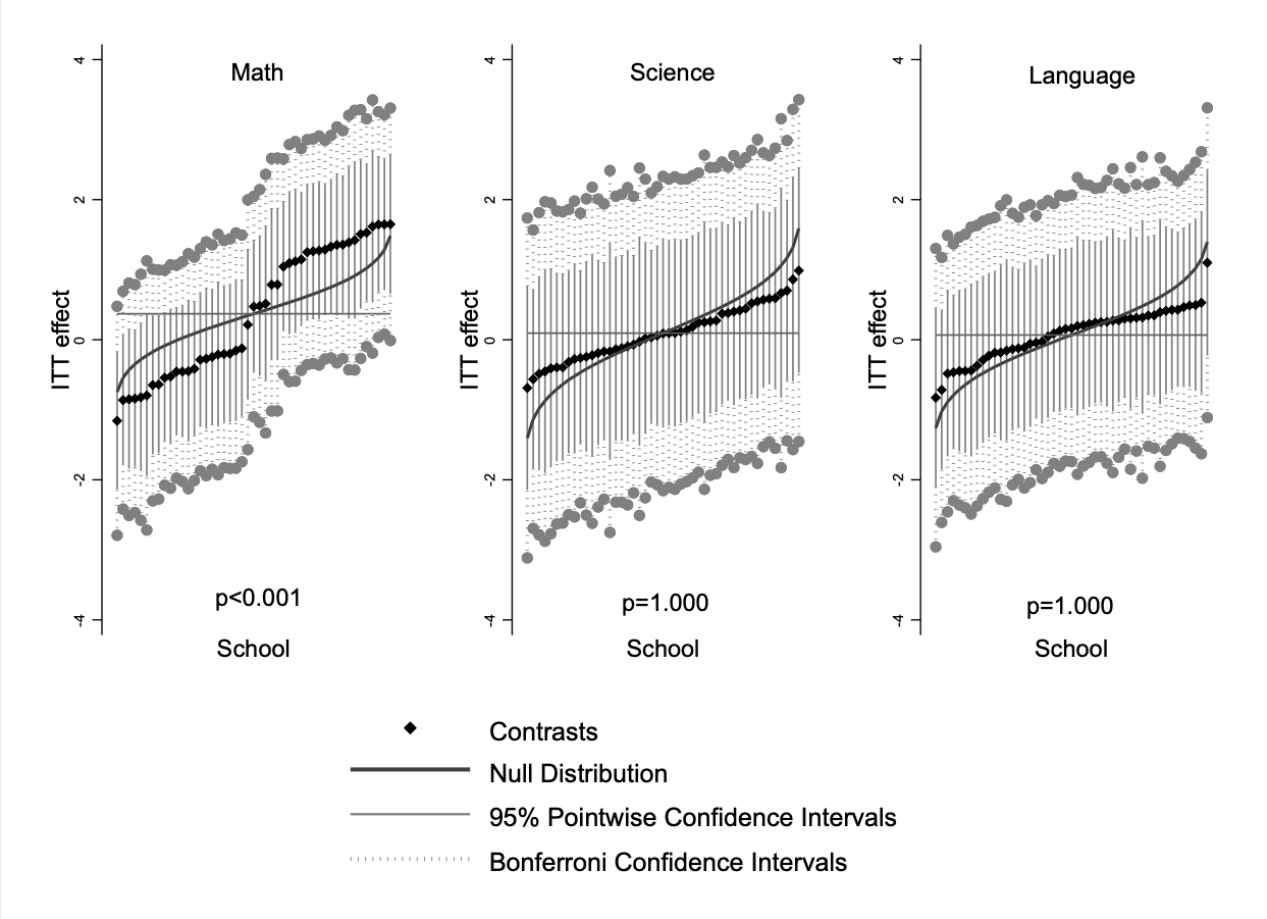
Notes: The figure provides a "caterpillar plot" (von Hippel and Bellows, 2018) of impacts on standardized scores by school at endline, about nine months after the rollout of the intervention (April 2018). Each black dot refers to the point estimate for a given school. Bonferroni confidence intervals adjust standard errors for multiple hypothesis testing. The black solid line shows the null distribution of "effects" that can be expected due to error. The p-values are for a test of the null hypothesis of no heterogeneity.

Figure A.7: Heterogeneous impact on standardized test scores by school (endline audit)



Notes: The figure provides a "caterpillar plot" (von Hippel and Bellows, 2018) of impacts on standardized scores by school at the endline "audit," about nine months after the rollout of the intervention (April 2018). Each black dot refers to the point estimate for a given school. Bonferroni confidence intervals adjust standard errors for multiple hypothesis testing. The black solid line shows the null distribution of "effects" that can be expected due to error. The p-values are for a test of the null hypothesis of no heterogeneity.

Figure A.8: Heterogeneous impact on standardized test scores by school (follow-up)



Notes: The figure provides a "caterpillar plot" (von Hippel and Bellows, 2018) of impacts on standardized scores by school at follow-up, about 11 months after the rollout of the intervention (June 2018). Each black dot refers to the point estimate for a given school. Bonferroni confidence intervals adjust standard errors for multiple hypothesis testing. The black solid line shows the null distribution of "effects" that can be expected due to error. The p-values are for a test of the null hypothesis of no heterogeneity.

Table A.1: Script adherence among SEI fellows (announced observations)

	(1) All fellows	(2) High-scoring	(3) Low-scoring
Share of SEI fellows who...			
...wrote on blackboard as indicated in the script	0.818	0.800	0.842
	[0.390]	[0.408]	[0.375]
...asked students questions in the script	0.750	0.800	0.684
	[0.438]	[0.408]	[0.478]
...used materials as indicated in the script	0.523	0.480	0.579
	[0.505]	[0.510]	[0.507]
...assigned students activities in the script	0.500	0.600	0.368
	[0.506]	[0.500]	[0.496]
...changed/rephrased parts of the script	0.341	0.280	0.421
	[0.479]	[0.458]	[0.507]
...added to or expanded on parts of the script	0.341	0.240	0.474
	[0.479]	[0.436]	[0.513]
...excluded parts of the script	0.182	0.160	0.211
	[0.390]	[0.374]	[0.419]
N (fellows)	50	28	22

Notes: This table displays the prevalence of script adherence among Science Education Initiative (SEI) fellows in treatment grades, based on announced observations about five months after the rollout of the intervention (December 2017). Standard deviations appear in brackets. High-scoring fellows are those who scored above the fellow-specific median on a test of content knowledge, instructional practices, and understanding of students' misconceptions (described in section 3); low-scoring fellows are those below that median.

Table A.2: Impact on allocation of time on task (announced observations)

	(1)	(2)	(3)	(4)	(5)
	PMC teachers		Impact on PMC teachers	SEI fellows	Difference between PMC teachers and SEI fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time...					
...reading aloud	0.049	0.000	-0.049***	0.014	-0.035**
	[0.122]	[0.000]	(0.013)	[0.041]	(0.014)
...on explanation/lecture	0.343	0.020	-0.323***	0.337	-0.007
	[0.208]	[0.061]	(0.022)	[0.191]	(0.030)
...on interactive demo	0.036	0.000	-0.036***	0.008	-0.028***
	[0.070]	[0.000]	(0.007)	[0.028]	(0.009)
...on question and answers	0.181	0.002	-0.179***	0.153	-0.028
	[0.153]	[0.014]	(0.016)	[0.126]	(0.021)
...on practice and drill	0.007	0.002	-0.005*	0.004	-0.003
	[0.030]	[0.014]	(0.003)	[0.020]	(0.005)
...on classwork	0.129	0.002	-0.127***	0.157	0.028
	[0.147]	[0.014]	(0.016)	[0.117]	(0.021)
...copying	0.032	0.000	-0.032***	0.057	0.025**
	[0.067]	[0.000]	(0.007)	[0.076]	(0.012)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Minutes of lesson time...					
...reading aloud	1.469	0.000	-1.469***	1.714	0.245
	[3.668]	[0.000]	(0.395)	[4.899]	(0.765)
...on explanation/lecture	10.281	2.400	-7.881***	40.408	30.067***
	[6.228]	[7.273]	(1.213)	[22.938]	(3.194)
...on interactive demo	1.094	0.000	-1.094***	0.980	-0.116
	[2.093]	[0.000]	(0.218)	[3.320]	(0.544)
...on question and answers	5.438	0.240	-5.198***	18.367	12.943***
	[4.592]	[1.697]	(0.511)	[15.120]	(2.009)
...on practice and drill	0.219	0.240	0.021	0.490	0.273
	[0.897]	[1.697]	(0.199)	[2.399]	(0.364)
...on classwork	3.875	0.240	-3.635***	18.857	14.969***
	[4.416]	[1.697]	(0.525)	[14.071]	(1.971)
...copying	0.969	0.000	-0.969***	6.857	5.892***
	[2.018]	[0.000]	(0.209)	[9.165]	(1.270)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of time on task of Pune Municipal Corporation (PMC) teachers in the control and treatment groups and of Science Education Initiative (SEI) fellows in the treatment group, measured in announced observations about five months after the rollout of the intervention (December 2017). All lessons taught by PMC teachers lasted 30 minutes and all lessons taught by SEI fellows lasted 120 minutes. Panel A displays results expressed as a proportion of lesson time and Panel B shows results in terms of minutes per lesson. Standard deviations appear in brackets and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Impact on allocation of time on classroom management (announced observations)

	(1)	(2)	(3)	(4)	(5)
	PMC teachers		Impact on PMC teachers	SEI fellows	Difference between PMC teachers and SEI fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time...					
...on instructions	0.011	0.006	-0.005	0.043	0.031***
	[0.035]	[0.024]	(0.005)	[0.076]	(0.010)
...on discipline	0.015	0.010	-0.005	0.016	0.002
	[0.038]	[0.036]	(0.006)	[0.051]	(0.008)
...on class mgmt. w/students	0.057	0.008	-0.049***	0.082	0.024*
	[0.084]	[0.027]	(0.011)	[0.075]	(0.014)
...on class mgmt. alone	0.055	0.184	0.129***	0.018	-0.037***
	[0.099]	[0.245]	(0.035)	[0.049]	(0.013)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Minutes of lesson time...					
...on instructions	0.344	0.720	0.376	5.143	4.799***
	[1.055]	[2.879]	(0.410)	[9.165]	(1.262)
...on discipline	0.438	1.200	0.762	1.959	1.523*
	[1.150]	[4.371]	(0.595)	[6.171]	(0.874)
...on class mgmt. w/students	1.719	0.960	-0.759	9.796	8.088***
	[2.529]	[3.289]	(0.582)	[9.058]	(1.312)
...on class mgmt. alone	1.656	22.080	20.424***	2.204	0.558
	[2.980]	[29.430]	(3.930)	[5.834]	(0.853)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of time on classroom management of Pune Municipal Corporation (PMC) teachers in the control and treatment groups and of Science Education Initiative (SEI) fellows in the treatment group, measured in announced observations about five months after the rollout of the intervention (December 2017). All lessons taught by PMC teachers lasted 30 minutes and all lessons taught by SEI fellows lasted 120 minutes. Panel A displays results expressed as a proportion of class time and Panel B shows results in terms of minutes per lesson. Standard deviations appear in brackets and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Impact on allocation of time off task (announced observations)

	(1)	(2)	(3)	(4)	(5)
	PMC teachers		Impact on PMC teachers	SEI fellows	Difference between PMC teachers and SEI fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time...					
...on social interaction w/students	0.007	0.002	-0.005	0.002	-0.005
	[0.030]	[0.014]	(0.004)	[0.014]	(0.004)
...on social interaction w/adults	0.015	0.016	0.001	0.006	-0.008
	[0.046]	[0.037]	(0.008)	[0.024]	(0.006)
...uninvolved	0.038	0.102	0.064*	0.018	-0.019
	[0.094]	[0.235]	(0.034)	[0.049]	(0.014)
...out of room	0.024	0.572	0.548***	0.069	0.046***
	[0.086]	[0.395]	(0.053)	[0.102]	(0.016)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Minutes of lesson time...					
...on social interaction w/students	0.219	0.240	0.021	0.245	0.030
	[0.897]	[1.697]	(0.263)	[1.714]	(0.261)
...on social interaction w/adults	0.438	1.920	1.482**	0.735	0.298
	[1.375]	[4.444]	(0.660)	[2.907]	(0.419)
...uninvolved	1.125	12.240	11.115***	2.204	1.084
	[2.829]	[28.220]	(3.925)	[5.834]	(0.927)
...out of the room	0.719	68.640	67.921***	8.327	7.628***
	[2.566]	[47.448]	(6.274)	[12.297]	(1.717)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of time off task of Pune Municipal Corporation (PMC) teachers in the control and treatment groups and of Science Education Initiative (SEI) fellows in the treatment group, measured in announced observations about five months after the rollout of the intervention (December 2017). All lessons taught by PMC teachers lasted 30 minutes and all lessons taught by SEI fellows lasted 120 minutes. Panel A displays results expressed as a proportion of class time and Panel B shows results in terms of minutes per lesson. Standard deviations appear in brackets and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: Impact on allocation of time engaging students (announced observations)

	(1)	(2)	(3)	(4)	(5)
	PMC teachers		Impact on PMC teachers	SEI fellows	Difference between PMC teachers and SEI fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time engaging...					
...no students	0.091	0.746	0.655***	0.106	0.016
	[0.143]	[0.296]	(0.042)	[0.121]	(0.022)
...one student	0.119	0.014	-0.105***	0.129	0.010
	[0.127]	[0.040]	(0.015)	[0.122]	(0.023)
...a small group of students	0.044	0.000	-0.044***	0.045	0.001
	[0.084]	[0.000]	(0.009)	[0.074]	(0.014)
...a large group of students	0.232	0.016	-0.216***	0.298	0.066
	[0.199]	[0.042]	(0.024)	[0.230]	(0.045)
...all students	0.451	0.024	-0.427***	0.404	-0.048
	[0.248]	[0.059]	(0.031)	[0.230]	(0.041)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Number of minutes engaging...					
...no students	2.719	89.520	86.801***	12.735	10.049***
	[4.289]	[35.566]	(4.775)	[14.576]	(2.024)
...one student	3.562	1.680	-1.883**	15.429	11.850***
	[3.803]	[4.855]	(0.765)	[14.697]	(2.142)
...a small group of students	1.312	0.000	-1.312***	5.388	4.070***
	[2.531]	[0.000]	(0.272)	[8.853]	(1.291)
...a large group of students	6.969	1.920	-5.049***	35.755	28.789***
	[5.976]	[5.062]	(0.952)	[27.549]	(4.302)
...all students	13.531	2.880	-10.651***	48.490	34.938***
	[7.425]	[7.093]	(1.472)	[27.600]	(4.203)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of time engaging students of Pune Municipal Corporation (PMC) teachers in the control and treatment groups and of Science Education Initiative (SEI) fellows in the treatment group, measured in announced observations about five months after the rollout of the intervention (December 2017). All lessons taught by PMC teachers lasted 30 minutes and all lessons taught by SEI fellows lasted 120 minutes. Panel A displays results expressed as a proportion of class time and Panel B shows results in terms of minutes per lesson. Standard deviations appear in brackets and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6: Impact on proportion-correct test scores (endline, endline audit, and follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.049***	0.067***	0.036***	0.038***	0.031**	0.035***
	(0.010)	(0.009)	(0.009)	(0.009)	(0.012)	(0.009)
Baseline score		0.584***		0.509***		0.520***
		(0.019)		(0.031)		(0.024)
N (students)	2524	2307	2524	2307	2524	2307
B. Endline audit						
Treatment	0.011	0.067***	0.092***	0.038***		
	(0.026)	(0.009)	(0.029)	(0.009)		
Baseline score		0.584***		0.509***		
		(0.019)		(0.031)		
N (students)	553	2307	553	2307		
C. Follow-up						
Treatment	0.058***	0.073***	0.019**	0.022**	0.006	0.008**
	(0.010)	(0.009)	(0.007)	(0.008)	(0.004)	(0.004)
Baseline score		0.609***		0.303***		0.186***
		(0.026)		(0.026)		(0.012)
N (students)	2369	2023	2369	2023	2369	2023

Notes: This table compares students' proportion-correct scores in the control and treatment groups based on assessments administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018) and at follow-up, about 11 months after the rollout of the intervention (June 2018). Scores are standardized to have mean zero and standard deviation one in the control group. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.7: Impact on IRT-scaled test scores (endline, endline audit, and follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.266***	0.347***	0.192***	0.192***	0.115**	0.159***
	(0.047)	(0.043)	(0.053)	(0.048)	(0.048)	(0.035)
Baseline score		0.703***		0.508***		0.598***
		(0.024)		(0.031)		(0.026)
N (students)	2504	2288	2497	2284	2432	2231
B. Endline audit						
Treatment	0.098	0.134	0.222***	0.236***		
	(0.085)	(0.085)	(0.067)	(0.063)		
Baseline score		0.140***		0.119***		
		(0.043)		(0.032)		
N (students)	553	551	553	551		
C. Follow-up						
Treatment	0.293***	0.353***	0.089*	0.106**	0.033	0.055
	(0.055)	(0.048)	(0.044)	(0.052)	(0.038)	(0.034)
Baseline score		0.672***		0.284***		0.354***
		(0.032)		(0.028)		(0.029)
N (students)	2270	1925	2271	1926	2267	1924

Notes: This table compares students' Item Response Theory (IRT)-scaled scores in the control and treatment groups based on assessments administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018) and at follow-up, about 11 months after the rollout of the intervention (June 2018). Scores are standardized to have mean zero and standard deviation one in the control group. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.8: Heterogeneous impact on standardized test scores by students' baseline score (endline, endline audit, follow-up)

	Math	Science	Language
	(1)	(2)	(3)
A. Endline			
Treatment	0.339***	0.214***	0.154***
	(0.048)	(0.050)	(0.038)
Baseline score	0.627***	0.463***	0.585***
	(0.030)	(0.037)	(0.029)
Treat × Baseline score	0.040	0.044	-0.063
	(0.038)	(0.040)	(0.043)
P-value of the sum	0.000	0.000	0.000
N (students)	2307	2307	2307
B. Endline audit			
Treatment	0.091	0.399***	
	(0.104)	(0.113)	
Baseline score	0.133*	0.206**	
	(0.078)	(0.078)	
Treat × Baseline score	0.014	-0.006	
	(0.121)	(0.105)	
P-value of the sum	0.129	0.021	
N (students)	551	551	
C. Follow-up			
Treatment	0.356***	0.138**	0.081**
	(0.046)	(0.052)	(0.036)
Baseline score	0.654***	0.291***	0.448***
	(0.039)	(0.035)	(0.037)
Treat × Baseline score	-0.006	0.061	-0.004
	(0.048)	(0.050)	(0.043)
P-value of the sum	0.000	0.000	0.000
N (students)	2023	2023	2023

Notes: This table shows the impact of the intervention on assessments of math, science, and language administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018), and at follow-up, about 11 months after the rollout of the intervention (June 2018) by students' baseline score. Endline scores are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.9: Heterogeneous impact on standardized test scores by students' socio-economic status (endline, endline audit, follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.261***	0.344***	0.191***	0.210***	0.113**	0.138***
	(0.051)	(0.048)	(0.052)	(0.050)	(0.052)	(0.041)
SES index	-0.030	-0.039*	-0.069*	-0.096***	-0.084**	-0.074**
	(0.033)	(0.021)	(0.037)	(0.032)	(0.037)	(0.029)
Treat × SES index	0.112**	0.056*	0.165***	0.144***	0.111*	0.057
	(0.045)	(0.032)	(0.053)	(0.045)	(0.058)	(0.046)
Baseline score		0.654***		0.502***		0.574***
		(0.023)		(0.032)		(0.026)
P-value of the sum	0.017	0.464	0.010	0.173	0.539	0.648
N (students)	2255	2255	2255	2255	2255	2255
B. Endline audit						
Treatment	0.059	0.098	0.392***	0.424***		
	(0.111)	(0.111)	(0.121)	(0.115)		
SES index	0.016	0.003	-0.082	-0.105		
	(0.080)	(0.079)	(0.067)	(0.066)		
Treat × SES index	-0.093	-0.091	0.106	0.104		
	(0.104)	(0.100)	(0.108)	(0.108)		
Baseline score		0.154**		0.224***		
		(0.064)		(0.063)		
P-value of the sum	0.210	0.117	0.751	0.985		
N (students)	532	532	532	532		
C. Follow-up						
Treatment	0.274***	0.355***	0.133***	0.152***	0.059	0.077**
	(0.052)	(0.045)	(0.049)	(0.053)	(0.048)	(0.036)
SES index	-0.001	-0.012	-0.048	-0.068**	-0.035	-0.032
	(0.033)	(0.024)	(0.030)	(0.029)	(0.032)	(0.026)
Treat × SES index	0.083*	0.036	0.138***	0.125***	0.089**	0.049
	(0.044)	(0.033)	(0.046)	(0.041)	(0.042)	(0.032)
Baseline score		0.658***		0.340***		0.451***
		(0.030)		(0.029)		(0.029)
P-value of the sum	0.031	0.405	0.013	0.057	0.139	0.555
N (students)	1972	1972	1972	1972	1972	1972

Notes: This table shows the impact of the intervention on assessments of math, science, and language administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018), and at follow-up, about 11 months after the rollout of the intervention (June 2018) by students' socio-economic status. Endline scores are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.10: Heterogeneous impact on standardized test scores by students' sex (endline, endline audit, follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.218**	0.267***	0.131	0.148**	0.025	0.095
	(0.083)	(0.065)	(0.080)	(0.070)	(0.074)	(0.064)
Female	-0.016	-0.004	0.064	0.053	0.161**	0.147***
	(0.067)	(0.047)	(0.087)	(0.073)	(0.067)	(0.053)
Treat × Female	0.085	0.145*	0.114	0.117	0.163*	0.080
	(0.104)	(0.081)	(0.114)	(0.105)	(0.089)	(0.088)
Baseline score		0.657***		0.501***		0.568***
		(0.022)		(0.032)		(0.028)
P-value of the sum	0.433	0.047	0.066	0.038	0.000	0.002
N (students)	2258	2258	2258	2258	2258	2258
B. Endline audit						
Treatment	-0.005	0.028	0.452***	0.487***		
	(0.136)	(0.141)	(0.124)	(0.126)		
Female	0.155	0.158	0.154	0.162		
	(0.171)	(0.167)	(0.131)	(0.124)		
Treat × Female	0.121	0.130	-0.142	-0.149		
	(0.219)	(0.221)	(0.202)	(0.199)		
Baseline score		0.155**		0.220***		
		(0.061)		(0.062)		
P-value of the sum	0.121	0.103	0.950	0.943		
N (students)	533	533	533	533		
C. Follow-up						
Treatment	0.262***	0.310***	0.096	0.104	0.052	0.114*
	(0.096)	(0.066)	(0.074)	(0.077)	(0.080)	(0.061)
Female	0.068	0.071	0.103	0.087	0.134*	0.115*
	(0.083)	(0.057)	(0.073)	(0.069)	(0.072)	(0.059)
Treat × Female	0.022	0.083	0.065	0.086	0.014	-0.066
	(0.122)	(0.081)	(0.099)	(0.103)	(0.103)	(0.086)
Baseline score		0.662***		0.340***		0.450***
		(0.030)		(0.028)		(0.029)
P-value of the sum	0.305	0.012	0.029	0.020	0.069	0.435
N (students)	1974	1974	1974	1974	1974	1974

Notes: This table shows the impact of the intervention on assessments of math, science, and language administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018), and at follow-up, about 11 months after the rollout of the intervention (June 2018) by students' sex. Endline scores are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.11: Heterogeneous impact on standardized test scores by students' caste (endline, endline audit, follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.337***	0.389***	0.223***	0.205***	0.108	0.077
	(0.055)	(0.045)	(0.067)	(0.065)	(0.066)	(0.067)
Scheduled caste/tribe	-0.111	-0.017	-0.101	-0.071	-0.132	-0.122**
	(0.078)	(0.047)	(0.084)	(0.065)	(0.081)	(0.060)
Treat × SC/ST	-0.106	-0.092	-0.018	0.022	0.079	0.160*
	(0.088)	(0.062)	(0.101)	(0.087)	(0.091)	(0.083)
Baseline score		0.647***		0.494***		0.559***
		(0.021)		(0.027)		(0.025)
P-value of the sum	0.003	0.061	0.094	0.405	0.538	0.602
N (students)	2236	2236	2236	2236	2236	2236
B. Endline audit						
Treatment	0.206	0.232	0.517***	0.523***		
	(0.165)	(0.165)	(0.155)	(0.154)		
Scheduled caste/tribe	0.019	0.043	0.048	0.038		
	(0.139)	(0.143)	(0.102)	(0.100)		
Treat × SC/ST	-0.274	-0.276	-0.188	-0.169		
	(0.197)	(0.198)	(0.178)	(0.181)		
Baseline score		0.129**		0.174***		
		(0.064)		(0.061)		
P-value of the sum	0.042	0.063	0.369	0.411		
N (students)	537	537	537	537		
C. Follow-up						
Treatment	0.307***	0.366***	0.178**	0.158**	0.136**	0.096*
	(0.055)	(0.051)	(0.068)	(0.065)	(0.056)	(0.053)
Scheduled caste/tribe	-0.082	0.013	-0.057	-0.048	-0.001	-0.009
	(0.071)	(0.043)	(0.063)	(0.057)	(0.064)	(0.061)
Treat × SC/ST	-0.025	-0.023	-0.090	-0.041	-0.090	-0.006
	(0.086)	(0.061)	(0.093)	(0.085)	(0.086)	(0.085)
Baseline score		0.665***		0.323***		0.448***
		(0.026)		(0.029)		(0.030)
P-value of the sum	0.202	0.866	0.037	0.153	0.173	0.780
N (students)	1955	1955	1955	1955	1955	1955

Notes: This table shows the impact of the intervention on assessments of math, science, and language administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018), and at follow-up, about 11 months after the rollout of the intervention (June 2018) by students' caste. Endline scores are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.12: Heterogeneous impact on standardized test scores by student-instructor sex match (endline, endline audit, follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.312***	0.367***	0.220***	0.242***	0.148**	0.161***
	(0.068)	(0.047)	(0.071)	(0.061)	(0.066)	(0.059)
Sex match	0.084	0.042	0.147	0.124	0.228***	0.137**
	(0.074)	(0.049)	(0.097)	(0.076)	(0.075)	(0.057)
Treat × Sex match	-0.037	0.029	0.014	-0.010	-0.022	0.003
	(0.101)	(0.087)	(0.122)	(0.109)	(0.116)	(0.105)
Baseline score		0.656***		0.505***		0.566***
		(0.021)		(0.030)		(0.028)
P-value of the sum	0.585	0.327	0.107	0.203	0.037	0.104
N (students)	2128	2128	2128	2128	2128	2128
B. Endline audit						
Treatment	-0.054	-0.016	0.466***	0.505***		
	(0.137)	(0.137)	(0.152)	(0.140)		
Sex match	-0.028	-0.032	0.146	0.165		
	(0.170)	(0.165)	(0.134)	(0.119)		
Treat × Sex match	0.344	0.346	-0.180	-0.207		
	(0.251)	(0.255)	(0.244)	(0.243)		
Baseline score		0.154**		0.230***		
		(0.065)		(0.063)		
P-value of the sum	0.145	0.145	0.872	0.846		
N (students)	506	506	506	506		
C. Follow-up						
Treatment	0.337***	0.378***	0.121**	0.128*	0.101	0.100**
	(0.081)	(0.055)	(0.059)	(0.064)	(0.062)	(0.048)
Sex match	0.159*	0.105*	0.103	0.076	0.150*	0.071
	(0.084)	(0.058)	(0.065)	(0.055)	(0.079)	(0.064)
Treat × Sex match	-0.068	0.012	0.089	0.094	-0.039	-0.012
	(0.127)	(0.093)	(0.094)	(0.093)	(0.102)	(0.079)
Baseline score		0.655***		0.342***		0.449***
		(0.030)		(0.030)		(0.028)
P-value of the sum	0.318	0.076	0.024	0.039	0.128	0.288
N (students)	1870	1870	1870	1870	1870	1870

Notes: This table shows the impact of the intervention on assessments of math, science, and language administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018), and at follow-up, about 11 months after the rollout of the intervention (June 2018) by whether the sex of the instructor matches the sex of the student. Endline scores are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.13: Heterogeneous impact on standardized test scores by instructors' assessment score (endline, endline audit, follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.255**	0.449***	0.162	0.211**	0.127	0.242***
	(0.101)	(0.064)	(0.107)	(0.091)	(0.123)	(0.082)
Instructor score	0.104*	0.068	0.117**	0.093*	0.039	0.021
	(0.058)	(0.046)	(0.053)	(0.050)	(0.075)	(0.052)
Treat × Instructor score	-0.083	-0.123**	-0.064	-0.073	-0.026	-0.079
	(0.089)	(0.053)	(0.098)	(0.074)	(0.105)	(0.074)
Baseline score		0.649***		0.486***		0.557***
		(0.021)		(0.029)		(0.025)
P-value of the sum	0.723	0.119	0.444	0.687	0.873	0.395
N (students)	2374	2174	2374	2174	2374	2174
B. Endline audit						
Treatment	0.099	0.167	0.295	0.365*		
	(0.236)	(0.242)	(0.219)	(0.214)		
Instructor score	-0.178*	-0.196**	0.014	-0.011		
	(0.097)	(0.095)	(0.096)	(0.094)		
Treat × Instructor score	0.119	0.123	0.046	0.037		
	(0.196)	(0.198)	(0.190)	(0.179)		
Baseline score		0.169***		0.212***		
		(0.061)		(0.063)		
P-value of the sum	0.704	0.641	0.689	0.849		
N (students)	523	521	523	521		
C. Follow-up						
Treatment	0.233**	0.368***	0.052	0.106	0.008	0.090
	(0.114)	(0.080)	(0.078)	(0.064)	(0.083)	(0.064)
Instructor score	0.118*	0.072	0.028	0.004	0.051	-0.004
	(0.064)	(0.046)	(0.036)	(0.038)	(0.043)	(0.032)
Treat × Instructor score	-0.059	-0.070	0.028	0.024	-0.004	0.001
	(0.101)	(0.069)	(0.067)	(0.058)	(0.072)	(0.057)
Baseline score		0.649***		0.325***		0.446***
		(0.029)		(0.030)		(0.027)
P-value of the sum	0.323	0.955	0.272	0.533	0.338	0.956
N (students)	2231	1909	2231	1909	2231	1909

Notes: This table shows the impact of the intervention on assessments of math, science, and language administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018), and at follow-up, about 11 months after the rollout of the intervention (June 2018) by instructors' standardized scores on an assessment of instructional practice, knowledge of student errors, and content knowledge in math (calculated as the first principal component of a principal-component analysis of the scores on all three domains, standardized with respect to instructors in the control group). Endline scores are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.14: Impact on standardized test scores among high-scoring teachers (endline, endline audit, follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	0.256***	0.390***	0.176**	0.200**	0.102	0.193***
	(0.069)	(0.060)	(0.085)	(0.079)	(0.091)	(0.054)
Baseline score		0.634***		0.460***		0.554***
		(0.032)		(0.035)		(0.028)
N (students)	1382	1276	1382	1276	1382	1276
B. Endline audit						
Treatment	0.252	0.331*	0.305*	0.337*		
	(0.175)	(0.175)	(0.176)	(0.172)		
Baseline score		0.214***		0.144**		
		(0.075)		(0.068)		
N (students)	308	307	308	307		
C. Follow-up						
Treatment	0.232**	0.371***	0.118	0.160**	0.039	0.127**
	(0.093)	(0.076)	(0.076)	(0.064)	(0.077)	(0.053)
Baseline score		0.675***		0.287***		0.440***
		(0.042)		(0.036)		(0.031)
N (students)	1312	1130	1312	1130	1312	1130

Notes: This table shows the impact of the intervention on assessments of math, science, and language administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018), and at follow-up, about 11 months after the rollout of the intervention (June 2018) among teachers who obtained a proportion-correct score between 60 and 80% on an assessment of instructional practice, knowledge of student errors, and content knowledge in math. Endline scores are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.15: Impact on student attitudes, mindsets, and aspirations (endline)

	(1) Control	(2) Treatment	(3) Difference	(4) N
A. Math and science				
Share of students who...				
...enjoy learning math	0.814	0.830	0.017	2,524
	[0.390]	[0.376]	(0.017)	
...enjoy learning science	0.740	0.773	0.033	2,524
	[0.439]	[0.419]	(0.020)	
...wish they did not have math	0.277	0.257	-0.019	2,524
	[0.448]	[0.437]	(0.026)	
...wish they did not have science	0.278	0.257	-0.020	2,524
	[0.448]	[0.437]	(0.028)	
...feel nervous about math	0.422	0.380	-0.041	2,524
	[0.494]	[0.486]	(0.027)	
...feel nervous about science	0.402	0.404	0.002	2,524
	[0.491]	[0.491]	(0.028)	
...find math useful for life	0.728	0.761	0.033*	2,524
	[0.445]	[0.427]	(0.020)	
...find science useful for life	0.721	0.725	0.005	2,524
	[0.449]	[0.446]	(0.020)	
...stop trying when math gets hard	0.234	0.262	0.029	2,524
	[0.424]	[0.440]	(0.025)	
...stop trying when science gets hard	0.193	0.234	0.041*	2,524
	[0.395]	[0.424]	(0.022)	
Composite index (std.)	-0.000	0.038	0.039	2,524
	[1.000]	[0.992]	(0.041)	
B. Intelligence				
Share of students who believe...				
...people cannot change their intelligence	0.538	0.519	-0.019	2,524
	[0.499]	[0.500]	(0.028)	
...people have a fixed amount of intelligence	0.461	0.447	-0.014	2,524
	[0.499]	[0.497]	(0.028)	
...only some people are people	0.439	0.424	-0.014	2,524
	[0.496]	[0.494]	(0.025)	
...boys are more intelligent than girls	0.385	0.365	-0.020	2,524
	[0.487]	[0.482]	(0.025)	
...boys are better at math and science	0.371	0.357	-0.014	2,524
	[0.483]	[0.479]	(0.029)	
Composite index (std.)	-0.000	-0.052	-0.050	2,524
	[1.000]	[1.015]	(0.061)	
C. Aspirations				
Share of students who...				
...want to continue studying after high school	0.675	0.696	0.021	2,246
	[0.469]	[0.460]	(0.024)	
...want to study a STEM subject in high school	0.756	0.800	0.044	2,239
	[0.430]	[0.400]	(0.027)	
...want a STEM-related job	0.333	0.334	0.000	2,231
	[0.472]	[0.472]	(0.027)	
Composite index (std.)	-0.000	0.044	0.042	2,216
	[1.000]	[0.975]	(0.052)	

Notes: This table compares students' attitudes, midnsets, and aspirations in the control and treatment groups based on surveys administered at endline about nine months after the rollout of the intervention (April 2018). It shows the means and standard deviations for each group (columns 1-2) and tests for differences between groups including randomization-strata (i.e., grade) fixed effects (column 3). Standard deviations appear in brackets, and standard errors (clustered by school) appear in parentheses. The composite indexes are the first principal components of the variables in each panel, standardized with respect to the control group. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.16: Heterogeneous impact on student beliefs by students' sex (endline)

	STEM	Intelligence	Aspirations
	(1)	(2)	(3)
A. Endline			
Treatment	0.067	-0.017	0.115
	(0.067)	(0.068)	(0.076)
Female	0.080	-0.326***	0.220***
	(0.072)	(0.081)	(0.079)
Treat × Female	-0.010	-0.022	-0.058
	(0.091)	(0.112)	(0.096)
P-value of the sum	0.211	0.000	0.014
N (students)	2258	2258	2009

Notes: This table shows the impact of the intervention on surveys administered at endline about nine months after the rollout of the intervention (April 2018), by students' sex. Indexes are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.17: Heterogeneous impact on student beliefs by students' caste (endline)

	STEM	Intelligence	Aspirations
	(1)	(2)	(3)
A. Endline			
Treatment	0.060	-0.115*	0.148**
	(0.063)	(0.068)	(0.065)
Scheduled caste/tribe	-0.090	-0.112**	0.042
	(0.056)	(0.048)	(0.067)
Treat × SC/ST	0.009	0.122	-0.114
	(0.081)	(0.085)	(0.094)
P-value of the sum	0.158	0.882	0.248
N (students)	2236	2236	1989

Notes: This table shows the impact of the intervention on surveys administered at endline about nine months after the rollout of the intervention (April 2018), by students' socio-economic status. Indexes are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.18: Heterogeneous impact on student beliefs by students' socio-economic status (endline)

	STEM	Intelligence	Aspirations
	(1)	(2)	(3)
A. Endline			
Treatment	0.066	-0.032	0.084
	(0.043)	(0.065)	(0.052)
SES index	0.007	0.040	0.038
	(0.039)	(0.032)	(0.035)
Treat × SES index	-0.014	-0.082*	0.025
	(0.048)	(0.045)	(0.048)
P-value of the sum	0.806	0.176	0.102
N (students)	2255	2255	2006

Notes: This table shows the impact of the intervention on surveys administered at endline about nine months after the rollout of the intervention (April 2018), by students' socio-economic status. Indexes are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.19: Heterogeneous impact on student beliefs by baseline score (endline)

	STEM	Intelligence	Aspirations
	(1)	(2)	(3)
A. Endline			
Treatment	0.063	-0.050	0.094*
	(0.042)	(0.064)	(0.055)
Baseline score	0.034	-0.042*	0.111***
	(0.022)	(0.023)	(0.023)
Treat × Baseline	-0.046	0.018	0.008
	(0.028)	(0.037)	(0.027)
P-value of the sum	0.568	0.396	0.000
N (students)	2307	2307	2052

Notes: This table shows the impact of the intervention on surveys administered at endline about nine months after the rollout of the intervention (April 2018), by students' socio-economic status. Indexes are standardized to have mean zero and standard deviation one in the control group. The p-value of the sum is from a test that the sum of the main effect of treatment and its interaction is equal to zero. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.20: Impact on student attendance (unannounced visits, school registers, and endline)

	(1) Control	(2) Treatment	(3) Difference	(4) N
A. Unannounced visits				
Share of students observed at school	0.787	0.789	0.002	2,442
	[0.410]	[0.408]	(0.028)	
B. School registers				
Share of students marked as...				
...absent 0 days this week	0.551	0.544	-0.008	2,619
	[0.498]	[0.498]	(0.025)	
...absent 1-2 days this week	0.274	0.271	-0.003	2,619
	[0.446]	[0.445]	(0.022)	
...absent 3-4 days this week	0.074	0.083	0.009	2,619
	[0.261]	[0.276]	(0.012)	
...absent 5+ days this week	0.101	0.102	0.001	2,619
	[0.301]	[0.303]	(0.017)	
C. Endline student survey				
Share of students who reported being...				
...late 0 days this week	0.401	0.388	-0.014	2,219
	[0.490]	[0.487]	(0.030)	
...late 1-2 days this week	0.299	0.337	0.038**	2,219
	[0.458]	[0.473]	(0.018)	
...late 3-4 days this week	0.165	0.140	-0.025	2,219
	[0.372]	[0.348]	(0.020)	
...late 5+ days this week	0.135	0.135	0.001	2,219
	[0.341]	[0.342]	(0.018)	
...absent 0 days this week	0.410	0.412	0.002	2,222
	[0.492]	[0.492]	(0.031)	
...absent 1-2 days this week	0.334	0.308	-0.026	2,222
	[0.472]	[0.462]	(0.022)	
...absent 3-4 days this week	0.128	0.148	0.021	2,222
	[0.334]	[0.356]	(0.017)	
...absent 5+ days this week	0.129	0.133	0.003	2,222
	[0.335]	[0.339]	(0.017)	

Notes: This table compares the attendance of students in the control and treatment groups based on various measures collected during unannounced visits to schools about seven months after the rollout of the intervention (February 2018) and at endline two months later (April 2018). It shows the means and standard deviations for each group (columns 1-2) and tests for differences between groups including randomization-strata (i.e., grade) fixed effects (column 3). Standard deviations appear in brackets, and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.21: Impact on student demand for tuition (endline)

	(1) Control	(2) Treatment	(3) Difference	(4) N
A. Subjects				
Share of students who attend tuition...				
...in math	0.323	0.352	0.030	2,195
	[0.468]	[0.478]	(0.026)	
...in science	0.256	0.286	0.029	2,173
	[0.437]	[0.452]	(0.026)	
...in language	0.238	0.266	0.029	2,151
	[0.426]	[0.442]	(0.028)	
...in English	0.318	0.345	0.027	2,196
	[0.466]	[0.476]	(0.028)	
B. Duration				
Share of students who attend tuition...				
...less than 2 hours per week	0.036	0.038	0.002	2,237
	[0.186]	[0.192]	(0.010)	
...2-4 hours per week	0.097	0.099	0.003	2,237
	[0.296]	[0.299]	(0.015)	
...4-6 hours per week	0.048	0.046	-0.001	2,237
	[0.215]	[0.210]	(0.011)	
...more than 6 hours per week	0.216	0.252	0.036*	2,237
	[0.412]	[0.434]	(0.020)	

Notes: This table compares students' demand for private tuition in the control and treatment groups based on surveys administered at endline about nine months after the rollout of the intervention (April 2018). It shows the means and standard deviations for each group (columns 1-2) and tests for differences between groups including randomization-strata (i.e., grade) fixed effects (column 3). Standard deviations appear in brackets, and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.22: Impact on class materials (announced observations)

	(1) Control	(2) Treatment	(3) Col. (2)-(1)	(4) N
Class has blackboard	0.980	1.000	0.020	150
	[0.141]	[0.000]	(0.020)	
Class has whiteboard	0.020	0.000	-0.020	150
	[0.141]	[0.000]	(0.020)	
Class has chalk/markers	0.980	1.000	0.020	150
	[0.141]	[0.000]	(0.019)	
Class has textbook for teachers	0.820	0.300	-0.520***	150
	[0.388]	[0.461]	(0.083)	
Class has textbook for students	0.720	0.260	-0.460***	150
	[0.454]	[0.441]	(0.098)	
Class has laptop	0.060	0.040	-0.020	150
	[0.240]	[0.197]	(0.045)	
Class has digital whiteboard	0.040	0.040	0.000	150
	[0.198]	[0.197]	(0.004)	
Class has LCD projector	0.080	0.060	-0.020	150
	[0.274]	[0.239]	(0.036)	
Class has TV	0.160	0.060	-0.100	150
	[0.370]	[0.239]	(0.066)	
Class has science/math equipment	0.300	0.160	-0.140*	150
	[0.463]	[0.368]	(0.081)	
Class has maps	0.240	0.120	-0.120	150
	[0.431]	[0.327]	(0.078)	
Class has charts/poster	0.760	0.780	0.020	150
	[0.431]	[0.416]	(0.080)	
Class has toys/games	0.080	0.040	-0.040	150
	[0.274]	[0.197]	(0.040)	
Composite index (std.)	0.000	-0.126	-0.126	150
	[1.000]	[0.711]	(0.101)	

Notes: This table compares the availability of materials in control and treatment classes, based on announced observations about five months after the rollout of the intervention (December 2017). Standard deviations appear in brackets, and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.23: Impact on variability of standardized test scores (endline, endline audit, and follow-up)

	Math		Science		Language	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Endline						
Treatment	-0.001	0.014	0.015	0.030	-0.053*	-0.061*
	(0.035)	(0.034)	(0.040)	(0.038)	(0.030)	(0.033)
Baseline score		0.287**		0.330**		0.419**
		(0.122)		(0.163)		(0.207)
N (students)	93	93	93	93	93	93
B. Endline audit						
Treatment	-0.119	-0.114	0.015	0.017		
	(0.075)	(0.069)	(0.064)	(0.066)		
Baseline score		0.080		0.039		
		(0.259)		(0.240)		
N (students)	91	91	91	91		
C. Follow-up						
Treatment	0.006	0.032	0.129***	0.136***	-0.000	-0.007
	(0.039)	(0.038)	(0.035)	(0.037)	(0.035)	(0.034)
Baseline score		0.432***		0.157		0.437***
		(0.131)		(0.189)		(0.153)
N (students)	94	94	94	94	94	94

Notes: This table compares the variability in students' standardized scores in the control and treatment groups based on assessments administered at endline and endline "audit", about nine months after the rollout of the intervention (April 2018) and at follow-up, about 11 months after the rollout of the intervention (June 2018). Scores are standardized to have mean zero and standard deviation one in the control group. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.