Leveraging Complementarities between Teachers' Content Knowledge and Pedagogical Supports: Experimental Evidence from India*

Alejandro J. Ganimian†
New York University

Isaac M. Mbiti‡
University of Notre Dame

Abhilash Mishra§
Equitech Futures/
University of Chicago

August 17, 2026

Abstract

We evaluate whether pairing strong subject-matter knowledge with pedagogical supports can improve learning in public primary schools. In Pune, India, we randomly assigned grades within 48 schools to an intervention that recruited college students in science, technology, engineering, and math ("fellows"), supported them with training, coaching, and scripts, and had them partly replace math and science teachers. Despite less education and experience, fellows scored 1.56 standard deviations (SDs) higher than teachers on an assessment of content knowledge and pedagogy. The intervention raised achievement after one year by 0.32 SDs in math, 0.23 SDs in science, and 0.16 SDs in language; math and science gains persisted three months later. Effects extended beyond potentially coached items. Fellows also used stronger pedagogical practices, and those with more knowledge adhered more closely to substantive script components, suggesting that scripts complemented their expertise. The intervention was more cost-effective than roughly two thirds of rigorously evaluated education interventions worldwide. The large pool of eligible college students near urban schools suggests that the model could scale to other Indian cities.

JEL codes: C93, I21, I22, I25

Keywords: student achievement, math and science education, teaching effectiveness, professional development, coaching, lesson scripts, India.

---

*We thank the Abdul Latif Jameel Poverty Action Lab's (J-PAL) Post-Primary Education initiative, the University of Chicago's Kevin Xu Initiative on Science, Technology, and Global Development, and the Jacobs Foundation for the funding for this study. We thank Priya Pethe and Rahul Panda for making this study possible. We also thank Austin Dempewolff, Sharnic Djaker, Joshua Gilbert, Mustufa Patel, and especially Rashmi Menon for excellent research assistance. Felipe Barrera-Osorio, Andrew Ho, Dan Koretz, Adrienne Lucas, Luke Miratrix, Eric Taylor, Ludger Woessmann, and seminar participants at CESifo, NYU, UVA, Harvard, LSE, AEFP, NEUDC, and SREE provided comments on earlier drafts. We registered the study in the AEA Trial Registry as AEARCTR-0002386. Research protocols were approved by the Institutional Review Boards at UVA and the Institute for Financial Management and Research. All views are our own and not those of our institutions.

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

‡Professor of Poverty and Education, Keough School of Global Affairs, University of Notre Dame. imbiti@nd.edu.

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

1 Introduction

There is mounting 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 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 key role in producing the math and science 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 conventional approach to this challenge, professional development, has often been ineffective (Popova et al., 2022), possibly because it tries to improve pedagogy without addressing gaps in teacher content knowledge. Because content knowledge and pedagogy are complements, reforms must address both to raise student learning (Mbiti et al., 2019).

We evaluated a novel intervention in India that seeks to improve teacher capacity in math and science by simultaneously alleviating constraints on both content knowledge and pedagogy. The Science Education Initiative (SEI) recruits college students who are majoring in science, technology, engineering, and math (STEM) fields to teach for a year in public primary schools. These fellows are selected through a competitive process (involving exams, interviews, and demonstration lessons), they receive pre- and in-service training, scripts for every lesson, and periodic feedback from instructional coaches. They teach, on average, for six hours per week, replacing part of the lesson time allocated to math and science otherwise taught by teachers.[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 knowledge may be even more pronounced in developing settings. Primary-school teachers in Bihar and Uttar Pradesh answered only 47% of mathematics items drawn from the curriculum they taught correctly (Dundar et al., 2014). Two thirds of grade 4 teachers in seven Sub-Saharan African countries could not solve an algebra problem intended for their students (Bold et al., 2017). Other international studies also document gaps in teachers' content knowledge (Schmidt et al., 2007; Tatto et al., 2012).

[5] Study schools did not have timetables, so we cannot calculate the share of instructional time that shifted from teachers to fellows. If we assume a four-hour school day, six hours/week represents 30% of class time.

They receive a stipend of USD 4 per class, which is about 8% of the median teacher salary. We conducted a randomized evaluation of this program in the city of Pune, which is in the state of Maharashtra. We randomly assigned 48 schools to a fellow in either grade 5 or 6.[6]

Fellows had less schooling than teachers, no teaching degrees, and had never taught before. Yet, they outscored treatment teachers by 1.56 SDs ($p < 0.01$) on an assessment of math and science knowledge, optimal approaches to teach concepts and procedures, and understanding of student errors.[7] They also scored 0.60 SDs higher on an index of practices that promoted discussion, 0.58 SDs higher on an index of practices that fostered practice, and 0.35 SDs higher on an index of practices that cultivated a supportive classroom climate ($p < 0.01$ for all).

Consistent with these improvements in instruction, the intervention had moderate-to-large positive impacts on student achievement. Treatment students outperformed their control peers by 0.32 SDs in math and by 0.23 SDs in science ($p < 0.01$ for both). At endline, we found that the organization had used items from the baseline test in instructional materials. Effects on student learning, however, cannot be solely explained by coaching on such items. Treatment students outperformed their control peers on items that were introduced at endline (and thus, were not susceptible to coaching).[8] They also fared better on a contemporaneous test administered to a random sub-sample, even if it was less well aligned with the curriculum.[9] And perhaps most compellingly, they still scored 0.33 SDs better in math ($p < 0.01$) and 0.09 SDs in science ($p < 0.1$) at the start of the next school year, three months after the program.[10] The endline and follow-up gains in math and science achievement were broad-based, with the treatment distributions stochastically dominating the control distributions in both rounds.[11]

In fact, although fellows only taught math and science, the intervention had spillover effects on language test scores of 0.16 SDs at endline ($p < 0.01$). These effects show that impacts in math and science did not come at the expense of other subjects and offer further evidence against item coaching driving effects, as these items could not have been subject to coaching.

We find limited evidence that the intervention increased students' motivation and effort. It had null effects on attitudes towards math and science, views about the malleability of

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

[7] This test was administered at endline, so we cannot distinguish selection from treatment effects.

[8] The impacts on these items were 0.23 SDs in math and 0.14 SDs in science ($p < 0.01$). We also estimated item-specific treatment effects and found positive effects on non-repeated items in both subjects and rounds.

[9] The effects on this test were 0.1 SDs in math (non-significant) and 0.24 SDs in science ($p < 0.01$).

[10] For items first introduced at follow-up, the math impact was 0.22 SDs ($p < 0.01$). For science, the estimate was positive but not statistically significant (0.03 SDs).

[11] Control classrooms were more likely than treatment classrooms to have textbooks for teachers (52 pp.) and students (46 pp.; $p < 0.01$ for both), as well as math or science equipment (14 pp., $p < 0.1$). This pattern suggests compensatory behavior by principals and, if so, implies that our estimates understate the effects that would have arisen had resources been equally distributed across treatment and control classrooms.

intelligence or math and science skills, and aspirations to pursue STEM courses or careers.[12] Treatment students were no more likely than their control peers to attend school as measured by unannounced visits, attendance registers, or self-reports. We only find a 4 pp. ($p < 0.05$) increase in the share of students reporting being late 1-2 days in the week before the endline. They were no more likely to attend tuition in any subject or spend more time on tuition; we only find a marginally significant 4-pp. difference in spending 6 or more hours per week. They attempted 2 pp. more items in math ($p < 0.1$) and 1 pp. more in language ($p < 0.05$) in the endline tests, but there are no differences in test-taking persistence in the follow-up tests. These results suggest fellows did not raise learning just by getting students to work harder.[13]

The intervention meaningfully changed the instruction that students received at school. Treatment teachers were no more likely to attend school or to arrive on time than their control peers, and fellows did not have higher attendance or punctuality rates than treatment teachers. Yet, only 5% of present control teachers and 56% of present fellows were found in the classroom during unannounced visits at the start of the school day. Considering that control students were only taught by their teacher and treatment students were taught by their teacher or fellow, the latter were 38 pp. more likely ($p < 0.01$) to start receiving instruction on time. Further, during classroom observations, fellows scored 0.60 SDs higher than control teachers on an index of instructional practices promoting discussion, 0.58 SDs higher on an index of fostering practice, and 0.35 SDs higher on cultivating a supportive climate ($p < 0.01$ for all). Together, these results indicate that treatment students received more and better teaching.

Our study design allows us to estimate the combined effect of a bundle of four potentially complementary inputs: recruiting fellows, providing them with pre- and in-service training, scripting their instruction, and shifting some math and science lessons from teachers to fellows. 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 support that leads 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.[14]

We wanted to better understand, however, how scripts complemented fellows' knowledge, both because scripts were arguably the most important intensive support fellows received[15]

[12] We can rule out effects larger than 0.07 SDs on mindsets, and while effects on the other two indices are less precise, we find near-zero impacts on all indicators.

[13] Consistent with this interpretation, we find no evidence of role-model effects. Impacts on achievement, motivation, and effort did not vary with fellow-student sex match at endline or follow-up.

[14] Further, several prior studies have already evaluated some of these components on their own, including: pre- and in-service teacher training (Yoshikawa et al., 2015; Loyalka et al., 2019; Albornoz et al., 2020), coaching (Cilliers et al., 2019, 2022), and lesson scripting (Piper et al., 2014; Gray-Lobe et al., 2022a).

[15] Pre-service training only lasted five weeks, in-service training occurred once a week, and the frequency of coaching was variable. By contrast, fellows had a script for every lesson that they taught during the fellowship.

and because they provided precisely what the fellows lacked: clear guidance on how to teach. We find some evidence for the complementarity between fellow knowledge and script pedagogy. First, fellows used the scripts: 84% wrote on the blackboard, 79% asked students questions, 58% used materials as indicated, and 52% assigned scripted activities. Only 34% changed or rephrased script parts, 34% added to them, and 18% excluded them. Second, higher-scoring fellows followed more substantive script components and altered scripts less often, which suggests that scripts were complements—rather than substitutes—to fellows' expertise.[16] These results indicate scripts played an important role in the intervention. Consistent with these findings, we cannot reject the null hypothesis that impacts were the same across sites.

We contribute to global evidence on the importance of teachers' subject-matter expertise. Prior studies had documented strong correlations between teachers' content knowledge and students' achievement (Hill et al., 2005; Santibanez, 2006; Rockoff et al., 2011; Kane et al., 2013; Gitomer et al., 2014; Bold et al., 2017; Cruz et al., 2017; Hanushek et al., 2019), but causal evidence was limited to one quasi-experimental study (Metzler and Woessmann, 2012). We provide experimental evidence on an approach that recruits instructors with high content knowledge and equips them with pedagogical supports (training, lesson scripts, and coaching). Our estimates show that combining these two components can raise learning outcomes in public schools without requiring governments to replace their existing teacher workforce.

Our second main contribution is to growing research on the effectiveness of lesson scripts. Previous studies have focused on evaluating their impact as a *substitute* for teacher capacity: in settings where most teachers have low levels of education, scripts that tell them what to do at each stage of a lesson can ensure a minimum "floor" of instructional quality (Piper et al., 2018; Albornoz et al., 2020; Romero et al., 2020; Gray-Lobe et al., 2022a; Eble et al., 2021).[17] Our study illustrates how scripts can also act as *complements*: for teachers with subject-matter expertise and low pedagogical training and experience, they provide much-needed scaffolding. 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.

Finally, we contribute to the literature on how teachers respond to additional instructors. Our findings reinforce an emerging pattern: incumbents reduce effort when they and the new instructors perform the same role, but may increase it when their tasks are complementary (Duflo et al., 2011; Muralidharan and Sundararaman, 2013; Ganimian et al., 2023). In our setting, because fellows were assigned to the same sections as teachers—instead of to new, split classrooms—crowd-out did not reduce lesson time, showing that extra workers can improve instruction even when teachers offload work if they substitute for incumbents effectively.

[16] Above-median scoring fellows were 19 pp. more likely to ask scripted questions, 28 pp. more likely to assign scripted activities, 14 pp. less likely to change scripts, and 23 pp. less likely to add to them.

[17] 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).

2 Experiment

2.1 Context

We conducted our study in Pune city, the second-largest city in the Indian state of Maharashtra (after its capital city, Mumbai). According to the latest census of India (conducted in 2011), there are 3.1 million people in the city (MHA, 2012), rendering its population size comparable to that of countries such as Armenia, Mongolia, and Uruguay (UN, 2019). Its municipal school system is run by the Pune School Board (*Shikshan Mandal* or PSB) under the Pune Municipal Corporation (PMC). For broader context, in 2016–2017, Pune district, which includes the city and surrounding areas, had 3,473 primary-only schools (grades 1-5) and 1,941 primary schools with upper primary (grades 6-8), enrolling 834,354 students in total (NIEPA, 2017).[18]

Nearly all primary-school aged children in Pune district are enrolled in school: the net enrollment rate in upper primary is 91% (NIEPA, 2017). 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.[19] 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 in Pune district: 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 district lack basic math and science skills. According to a nationally representative student assessment administered by the central government, only 30% of fifth-graders in the district could use arithmetic for daily situations, 38% could identify equivalent fractions, and 34% could estimate a volume. Science results 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).

2.2 Sample

We selected 48 public and *Vidya Niketan* primary schools for this study.[20] We started with all 286 PSB-run primary schools. We excluded 118 schools away from the city center (because

[18]For reference, if Pune district 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).

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

[20]These schools are publicly funded and managed like regular public schools, but with more autonomy.

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),[21] 59 schools where the PSB or other non-profits were conducting other programs (because we wanted to estimate 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.[22] This randomization strategy, pioneered by Banerjee et al. (2007), seeks to minimize the risk of differential attrition (i.e., schools without any intervention dropping out 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 separate lotteries—one per preference set (e.g., one lottery for neighborhood 1, morning shift, English medium schools).

Table A.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%), suggesting that the schools where SEI places its fellows serve relatively low-income families. Yet, about one in three attend math tuition and one in four attend science tuition.[23]

We find no systematic differences in students across experimental groups. Only one of the 20 covariates we test (math achievement) differs ($p < 0.1$), which is what one would expect by chance. A joint F-test from a regression of treatment status on all listed covariates rejects

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

[22]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).

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

balance ($p = 0.013$), but this is driven by math achievement: excluding it yields $p = 0.159$. Yet, any resulting bias would attenuate, rather than inflate, our estimated treatment effects.

2.4 Attrition

We see no difference in attrition across control and treatment groups at endline or follow-up (Table A.2 in Appendix A). We also see no differences in the composition of the sample across experimental groups along students' characteristics with all interactions between treatment assignment and these characteristics statistically being insignificant in a model for attrition. Further, we see no evidence of differential attrition between control and treatment groups on observed baseline characteristics according to a joint test of significance across all interactions.

2.5 Intervention

The program we studied is the Science Education Initiative (SEI) undergraduate fellowship.[24] SEI is a Pune-based non-profit organization dedicated to improving math and science learning, and this fellowship was its flagship program. Since 2014, it has placed 200 fellows in 110 classes. It recruits, trains, and supports college majors 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;[25] (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.[26]

Fellows differ from teachers in background, education, experience, and knowledge (Table 1). 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.[27]

[24]SEI also has a fellowship for college graduates, which we did not evaluate in this study.

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

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

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

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.

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 on the blackboard, the activities to complete, and the questions to ask students.

Most fellows wrote on the board and asked students questions as specified in 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.3).[28] Interestingly, higher-scoring fellows asked more scripted questions, assigned more scripted activities, and made fewer changes, additions, or exclusions to 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 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.[29]

Fellows reported high satisfaction with the fellowship. In the endline survey, 92% agreed that the teaching method was very good, 96% agreed that the program provided quality instruction, and 98% agreed that it improved science and math learning (Table A.4). These positive responses were not accompanied by widespread implementation concerns: only 27% reported challenges with teachers, 20% reported that students did not respect them, and 12% reported challenges with school leadership. Fellows also expressed interest in continuing along this path, with 92% saying they would choose the fellowship again and 82% reporting interest in a teaching career after their experience in SEI.

3 Data

We conducted four data-collection rounds: a baseline at the start of the school year (to check the comparability of experimental groups, increase the precision of our impact estimates, and test for heterogeneous effects); a midline during the year (to identify potential impact mechanisms); an endline at the end of the year (to estimate the impact of the intervention);

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

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

and a follow-up at the start of the following school year (to check whether effects persisted). We list the dates, instruments administered, and participation rates per round in Table A.5.

3.1 Instructor surveys and assessments

We administered a survey and an assessment to teachers and fellows to describe how the intervention shifted the composition of instructors for some of the math and science lessons. We administered both instruments at endline, so their results may partly reflect the impact of the program on instructors (e.g., lesson scripts may have affected fellows' pedagogical skills).

We decided which domains to assess based on previous tests for educators (e.g., Hill et al., 2005; Bhattacharjea et al., 2011; Gitomer et al., 2014; Bold et al., 2017; World Bank, 2017). We assessed instructors' knowledge of the math and science curriculum for grades 5 and 6, the instructional practices they considered optimal to teach concepts and processes, and their understanding of the misconceptions that may explain mistakes that students frequently make. The test had 36 multiple-choice items from both domestic and international instruments.[30]

3.2 Student assessments

We administered assessments of math and science to measure the impact of the intervention. We also administered assessments of language to explore potential spillovers from the program. We assessed all three subjects on all data-collection rounds (baseline, endline, and follow-up). We decided which topics and skills to assess in each subject, and the share of items for each topic and skill, based on the frameworks for two international assessments (IEA, 2015, 2017). We tested three topics in math (numbers, geometry and measurement, data) and science (life, physical, and earth sciences) and three skills in both (knowledge, application, and reasoning). In language, we tested vocabulary, grammar, and reading, and whether students could retrieve explicit information, make inferences, and interpret and integrate ideas from written texts. Each test had 30 multiple-choice items, with repeated items across rounds for linking results. We scale test scores using a two-parameter logistic Item Response Theory (IRT) model and standardize them with respect to the control group within each subject and round.[31] We used items from domestic and international assessments and prior impact evaluations from a wide range of difficulty levels to prevent students answering none or all questions correctly.[32]

[30]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.

[31]We obtained the same substantive results using standardized proportion-correct scores.

[32]Figures A.2 show the distributions of proportion-correct scores on these assessments. As the figure shows, we see little evidence of floor or ceiling effects.

During the endline, we discovered that some of the items from our baseline assessments had been used in SEI's instructional materials, even if we had instructed them not to do so.[33] We identified 15 items in the program's practice tests, but we cannot rule out that more items were featured in other instructional materials. We address this challenge in five different ways. First, we find positive impacts on both the items from the baseline, which were susceptible to coaching, and those introduced at endline or follow-up, which could not have been coached. Second, following Gilbert et al. (2025), we estimate item-specific treatment effects, to show that many of the items introduced at endline and follow-up show positive treatment effects. Third, we find spillover effects on language, a subject that was not taught by the fellowship. Fourth, we show effects on an "audit" test of math and science administered to a sub-sample at endline, which was composed entirely of new items, and observe effects on this assessment.[34] Lastly, we document effects on a follow-up test administered three months after the end of the intervention, when any temporary coaching effects would have presumably attenuated.

3.3 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).

3.4 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 fellows to 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.[35]

3.5 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 fellows to teachers in

[33] Due to time constraints at the start of our study, we asked fellows to administer our baseline assessments (we administered all other rounds). We asked them not to review the assessments or take any copies home.

[34] We sourced this test from a contemporaneous evaluation in Kenya (Gray-Lobe et al., 2022b). These assessments were less closely aligned with the math and science curriculum for grades 5 and 6 in Pune, so we do not expect to see effects of similar magnitude to those in our endline and follow-up tests.

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

control and treatment classes.[36] We announced our observations because we were interested in how teachers used lesson time when they attended.[37] We also collected data on whether fellows and control teachers engaged in certain practices during the lesson.

4 Empirical strategy

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

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

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 ($t = 2$); $r(g)$ 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. X_{igs} is a vector of baseline covariates selected separately for each outcome using LASSO (Bruhn and McKenzie, 2009); the candidate set includes the baseline value of the outcome (when available) together with the student characteristics listed in Table A.1. 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, with 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. The endline scores of treatment students were 0.32 SDs higher in math and 0.23 SDs higher in science than those of their control peers ($p < 0.01$; Table 2, panel A).[38] This is true for all topics and all but one of the skills assessed across both subjects (Table A.7, panels A and B).

These effects persisted until the beginning of the following school year, three months after the intervention had ended. The follow-up scores of treatment students were 0.33 SDs higher

[36] 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).

[37] As stated above, schools in Pune do not have timetables. However, based on conversations with principals, we observed 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.

[38] In percent-correct units, these effects are 5.5 pp. in math and 4 pp. in science (Table A.6).

in math ($p < 0.01$) and 0.09 SDs higher in science ($p < 0.1$) than those of control students (Table 2, panel C), with positive impacts on all topics and skills (Table A.7, panels A and B).

These effects cannot be explained by fellows coaching students on the items from baseline. 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. We show that improvements on these items cannot fully account for the learning gains of treatment students in three ways. First, we estimate effects separately for items that were repeated from baseline to endline (and were susceptible to coaching) and for items that were introduced at endline or follow-up. Effects on the former are larger than those on the latter in both rounds, and the difference between them is statistically significant, but impacts on non-repeated items are positive and statistically significant in both rounds in math (0.23 SDs at endline and 0.22 SDs at follow-up, $p < 0.01$) and at endline in science (0.14 SDs at endline, $p < 0.01$; Table 2, panels A and C). Second, following Gilbert et al. (2025), we estimate item-specific treatment effects and find positive effects on non-repeated items in both subjects and rounds (Figure A.3, panels A-B). Third, treatment students also outperformed control students on an "audit" test from a study in primary schools in Kenya (Gray-Lobe et al., 2022a) that drew on different items, even if its content was better aligned with the curriculum in science than in math (Table 2, panel B).

The intervention had spillover effects on language. The endline scores of treatment students were 0.16 SDs higher than those of their control counterparts ($p < 0.01$; Table 2, panel A).[39] Impacts were positive for all topics and all but one of the skills assessed (Table A.7, panel C). Given that fellows did not teach language, spillovers likely stem from students applying what they learned in math and science to this subject. Consistent with this possibility, while the average difference between experimental groups in language was not statistically significant at follow-up (Table 2, panel A), impacts remained statistically significant on vocabulary and retrieving information—two domains often exercised in word problems (Table A.7, panel C).

Consistent with the intended effects of lesson scripting, a key component of the program, we find little evidence of heterogeneous effects across schools and students. When we estimate school-specific ITT effects, we fail to reject the null hypothesis that effects were equal across schools for science in both rounds and for math at endline (Figure A.4). When we estimate quantile effects, treatment distributions first-order stochastically dominate control distributions, and when we estimate average treatment effects by baseline achievement, we find positive impacts across the distribution (Figure A.5).[40] Impacts are larger for students of high socio-economic status (SES) in science and language at endline, but the interaction between

[39]In percentage-point terms, the effect is 3.5 pp. (Table A.6).

[40]We also find no heterogeneous effects by baseline scores (Table A.8, col. 1).

treatment and SES only remains statistically significant for science at follow-up (Table A.8).[41] We find little heterogeneity by other student or teacher characteristics (Table A.9).

Differences in classroom resources suggest our achievement estimates may be conservative. Whenever resources (e.g., fellows) are (randomly) allocated to some classes and not others, principals may try to compensate 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., $p < 0.1$; Table A.20). 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).

5.2 Student motivation and effort

The intervention had null effects on students' self-reported attitudes, mindsets, or aspirations. Treatment students did not differ from control students on any of the three composite indices. The 95% confidence intervals rule out positive effects larger than 0.07 SDs on beliefs about the malleability of intelligence and math skills. We cannot rule out moderate effects on attitudes toward math and science (0.21 SDs) or on educational and STEM aspirations (0.14 SDs), but we find near-zero impacts for nearly all indicators in both indices (Table A.10, panels A-C). We also see almost no evidence of heterogeneity, except for a marginally significant interaction effect on attitudes towards math and science for students with higher SES (Table A.11).

We also find limited evidence that the program impacted three measures of students' effort: attendance at school, demand for private tuition, or persistence on assessments. We measured attendance in three ways: calculating the share of enrolled students who were present during unannounced visits, digitizing school attendance registers, and asking students how often they were late to or absent from school. We only observe a 4 pp. ($p < 0.05$) increase in the share of students who report being late 1-2 days in the week prior to the endline (Table A.12). Treatment students were no more likely than their control peers to attend tuition, but they were 4 pp. ($p < 0.10$) more likely to spend 6 or more hours on tuition per week (Table A.13). At endline, treatment students attempted 2 pp. more items in math ($p < 0.10$) and 1 pp. more items in language ($p < 0.05$), but they were no more likely to reach the last item in the test because there was no margin for improvement: 99% of control students did so (Table A.14).

These results suggest fellows did not raise learning by just getting students to work harder. Accordingly, we find little evidence of role-model effects. If students viewed fellows as role

[41] Using mother's education instead of the SES index produces a similar pattern: impacts are larger among students whose mothers have more education, although the estimates are less precise.

models, the intervention's effects should be larger for students assigned to a fellow of the same sex. Yet, achievement impacts did not vary with fellow-student sex match at endline, endline audit, or follow-up in any subject (Table A.9). The fellowship's effects on students' attitudes, mindsets, and aspirations did not vary with fellow-student sex match either (Table A.11).

6 Mechanisms

6.1 Instructor attendance, punctuality, and instructional time

The intervention did not increase instructors' attendance. Treatment teachers were no more likely than their control peers to attend school and fellows did not go to school more often than treatment teachers. This is partly because attendance was already high: 84% of control and 88% of treatment teachers were in school during unannounced visits (Table 3, panel A). The program did not increase punctuality either: treatment and control teachers arrived on time at similar rates and fellows were 30 pp. less likely to be at school at the start of the day than treatment teachers.[42] Unlike teachers, fellows were only expected to be at school when teaching, so their lower punctuality may simply reflect their different schedule.[43]

The intervention increased students' probability of having an instructor in the classroom. During unannounced visits, enumerators recorded where they found the present instructors. Conditional on being at the school, fellows were 51 pp. more likely ($p < 0.01$) to be first spotted inside the classroom than either treatment or control teachers (Table 3, panel B). Conversely, present treatment teachers were 52 pp. more likely ($p < 0.01$) to be first seen in the principal's office (where teachers in this setting prepare lessons and socialize with peers). Given that control students could only receive instruction from their teacher and treatment students from either their teacher or a fellow, treatment students were 38 pp. more likely ($p < 0.01$) than their control peers to have an instructor at the start of the day (panel C).

When they were teaching, fellows did not allocate time differently from control teachers. This is partly because the latter already spent over three-fourths (78%) of lesson time on task, 14% on class management, and 8% off task, so there was little margin for improvement (Table 4, panel A).[44] The same pattern emerges if we break down these three categories into

[42]This pattern of results rules out discouragement effects among control teachers for not receiving a fellow. Control teachers were also no more likely to report higher workload or lower self-efficacy (Table A.15).

[43]As discussed above, schools did not have timetables, so we measured punctuality by checking whether an instructor had arrived before the enumerator reached the school or before the school's start time.

[44]As discussed in section 2.5, teachers taught either math or science in 30-minute lessons; fellows taught both math and science in 120-minute lessons. Therefore, while we compare these groups on the share of lesson time for each activity type, we also report the number of minutes on each category in panel B of Table 4.

more granular subcategories.[45] When fellows were teaching, however, treatment teachers were 70 pp. less likely than their control peers to be on task and 61 pp. more likely to be off task ($p < 0.01$ for both), with most off-task time being out of the room (Table A.18, panel B). This crowding out suggests that fellows substituted for treatment teachers instead of reducing student-teacher ratios through co-teaching or receiving mentoring from teachers during class.

6.2 Instructor pedagogical practices

The intervention improved instructors' pedagogical practices, particularly those that promoted discussion and fostered practice. Fellows scored 0.60 SDs higher than control teachers on an index of practices promoting discussion and 0.58 SDs higher on an index of practices fostering practice ($p < 0.01$ for both; Table 5, panels A and B). Relative to control teachers, fellows were 16-29 pp. more likely to ask both closed and open questions, ask students to explain their answers, correct student answers, and allow students to ask questions. They were also 16 pp. more likely to assign homework and 28 pp. more likely to help individual students. Fellows scored 0.35 SDs higher on an index of practices cultivating a supportive classroom climate ($p < 0.01$; panel C), driven by their greater propensity to praise or encourage students. By contrast, fellows and control teachers did not differ significantly on an index of practices engaging students during the announced classroom observations (panel D).

6.3 Lesson scripts

Fellows adhered to the lesson scripts, but adherence varied across components. In announced observations, 84% wrote on the board as indicated in the script and 79% asked students the scripted questions (Table A.3, panel A). Smaller shares used the prescribed materials (58%) or assigned the scripted activities (52%). Only a minority modified the scripts: 34% changed or rephrased parts, 34% added to or expanded on them, and 18% excluded parts (panel B).

More knowledgeable fellows adhered more closely to the instructional components. Compared with fellows who scored below the median on the instructor assessment, those at or above the median were 19 pp. more likely to ask the scripted questions and 28 pp. more likely to assign the scripted activities (Table A.3). They were also 14 pp. less likely to change or rephrase parts of the script and 23 pp. less likely to add to or expand on them. The two groups were similarly likely to use the prescribed blackboard content and materials. These patterns suggest that scripts complemented fellows' expertise rather than substituting for it.

Consistent with this complementarity, we find little evidence that scripts reduced the importance of instructor knowledge. We estimated the relationship between instructors'

[45] Control teachers spent most lesson time lecturing and explaining (34%), asking and answering questions (18%), and assigning students classwork (13%; Table A.16). The figures for fellows were nearly identical. Fellows also resembled control teachers in their use of class management and off task time (Tables A.17-A.18).

scores on the assessment of content knowledge, instructional practice, and student errors and students' achievement at endline and follow-up separately for control and treatment groups.[46] If scripts reduced the importance of instructor knowledge, the association between this score and achievement should be weaker in treatment classrooms. At endline, this association was weaker for both math and science, but the difference was statistically significant only for math, and we find no similar pattern at the endline audit or follow-up (Table A.9).[47]

7 Discussion

7.1 Cost-effectiveness

The SEI fellowship was more cost-effective than most other rigorously evaluated education interventions in LMICs. In the 2017-2018 school year, it cost INR 3.66 million for 60 fellows serving 1,920 students, or INR 1,907 (USD 30) per student (Table 6). These costs imply that the program raised test scores by 1.06 SDs in math and 0.77 SDs in science for each USD 100 spent. In a global dataset of cost-effectiveness estimates (Angrist et al., 2025), the median intervention generated about 0.4 SDs per USD 100, so the fellowship ranks around the 66th percentile in math and 64th percentile in science of that distribution.[48] The fellowship also compares favorably in learning-adjusted years of schooling (LAYS). Dividing our cost-effectiveness estimates by the 0.8 SD benchmark in Angrist et al. (2020) suggests that the fellowship generated 1.33 LAYS in math and 0.97 LAYS in science per USD 100 spent. In the same global dataset, the median intervention generated 0.49 LAYS per USD 100.

7.2 Scale-up

It seems feasible to recruit more STEM college students to teach, at least in large urban areas near postsecondary institutions. In Pune city, where we conducted our study, we estimate that 31,000-41,000 such students attend colleges within 5 km of a school in our sample.[49] The corresponding pools include 101,000 students in Pune district, 1.05 million in Maharashtra, and 10.2 million across India (MoE, 2026). Even if only 5% of eligible college students were

[46] We use teachers' scores in control classrooms and fellows' scores in treatment classrooms (section 6.1).

[47] Restricting the sample to overlapping scores yields similar effects across specifications (Table A.19).

[48] These figures exclude fellows' opportunity costs. Counting those costs under alternative wage assumptions puts per-student costs at USD 29.6-43.8, implying 0.72-1.07 SDs in math and 0.53-0.78 SDs in science per USD 100 spent. The corresponding LAYS ranges are 0.9-1.34 in math and 0.66-0.98 in science per USD 100; all ranges remain above the global database medians.

[49] We estimate Pune district's pool by applying the national share of science and engineering/technology students to Maharashtra's higher-education enrollment and allocating the result according to Pune district's share of the state's colleges. For Pune city, we geocode colleges offering these fields and estimate enrollment within 5 km of a study school. The lower estimate treats colleges that we could not geocode as outside this radius; the upper estimate assumes that they are distributed like the colleges that we could geocode.

willing to teach, these pools would yield 1,550-2,050 potential fellows in Pune city, 5,050 in Pune district, 52,500 in Maharashtra, and 510,000 across India.[50] This would represent a sizable share of the existing teacher workforce: 9%-11% of primary or upper-primary school teachers in Pune city, 10% in Pune district and Maharashtra, and 7% across India.[51]

7.3 Subject-matter expertise

Our results also shed light on the relationship between teacher content knowledge and student achievement. Until recently, the value of the former had been largely inferred from correlations with the latter (Hill et al., 2005; Santibanez, 2006; Rockoff et al., 2011; Kane et al., 2013; Gitomer et al., 2014; Bold et al., 2017; Cruz et al., 2017; Hanushek et al., 2019). We know of only one (quasi-experimental) causal 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.[52] The fellows in our study received considerable pedagogical support (training, lesson scripts, coaching), so we cannot attribute their impact to their expertise. Yet, it would make little theoretical sense to hire fellows without such supports, and doing so would likely affect who applies to the fellowship, so we believe we are estimating the policy-relevant parameter.

7.4 Lesson scripts

Our study advances our understanding of how structured pedagogy can support teachers. Structured-pedagogy interventions—including structured curriculum units, scripted lessons delivered by para-teachers, and tablet-delivered lesson plans—have consistently improved learning in LMICs (Piper et al., 2014, 2018; Albornoz et al., 2020; Eble et al., 2021; Gray-Lobe et al., 2022a; Akyeampong et al., 2023; Piper and Dubeck, 2024). More guidance does not always yield greater impacts: simplified guides are associated with larger gains than fully scripted ones (Piper et al., 2018), and shortening guides and increasing teacher discretion improved achievement (Gray-Lobe et al., 2025). Yet, prior studies have largely evaluated whether structured pedagogy can *substitute* for low teacher capacity; ours is one of the first to show that lesson scripting can *complement* teachers with high levels of content knowledge.

[50] Available data do not measure Indian STEM college students' willingness to join a paid, part-time teaching fellowship. Among college students in Delhi and Kolkata, 64%-72% would consider teaching, but only 11%-24% of those interested preferred school teaching (Dastidar and Sikdar, 2015); 11.4% of rural youth ages 14-18 aspired to become teachers (ASER, 2023). We therefore conservatively assume that 5% would teach.

[51] Teacher counts come from UDISE+ 2023-2024 and include all school-management types (MoE, 2024). We define the relevant workforce in each geographic area as teachers who report teaching primary, upper-primary, or adjacent combinations of grades in the UDISE+ data used for these calculations.

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

Our results suggest that scripts can help high-content-knowledge instructors translate what they know into effective teaching. Fellows embraced the structure, and those with higher assessment scores followed its substantive components more closely rather than exercising greater discretion. They also used stronger pedagogical practices without devoting more time to instruction. Together, these findings suggest that scripts need not constrain knowledgeable instructors; they may help them apply their expertise consistently in the classroom.

7.5 Additional instructors

We also contribute to evidence on the effects of introducing additional instructors in LMICs. In Busia and Teso, Kenya, teachers hired on renewable one-year contracts to lead a section of a split first-grade classroom had lower absence and were more likely to be observed teaching than regular teachers, but the latter became less likely to teach as a result (Duflo et al., 2011). In Andhra Pradesh, India, a similar initiative for multigrade primary-school classes had the same effects, also increasing regular-teacher absence (Muralidharan and Sundararaman, 2013), suggesting that they offloaded some of their teaching responsibilities to their contract peers. Yet, in Tamil Nadu, India, facilitators hired on an 18-month contract exclusively to teach reduced the main worker's absence and made childcare centers more likely to open on time (Ganimian et al., 2023), indicating that the facilitator actually crowded in the worker's effort.

The effects that we observe on teacher effort, which resemble most closely those in Kenya and India, reinforce one emerging pattern in this literature. When extra instructors are hired to perform the same role as existing staff—as was the case in these three studies—some crowd-out (in attendance and/or effort, conditional on attendance) is likely. But when they are hired to perform complementary tasks—as was the case in Tamil Nadu, where the main worker could focus on health and nutrition—they can increase specialization and productivity. In our study, given that fellows were hired to teach the same sections as teachers (instead of splitting existing classrooms), crowd-out did not lead to a net reduction in instructional time.

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 with pedagogical support. We find that the introduction of these fellows results in moderate-to-large gains in 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., increases in instructors' likelihood to be in their class and in their use of positive pedagogical practices) and largely rule out confounders such as instructor attendance, student attendance, and private tuition.

Overall, our study is best regarded as an efficacy trial showing that a carefully supported alternative staffing model can improve instruction in settings where teacher content knowledge is a binding constraint. The promise of the model is likely greatest in urban areas with nearby colleges and implementing organizations that can recruit, select, train, and support fellows. Testing whether these conditions can be sustained at larger scale, and whether the model can work outside college clusters, remains an important area for future research.

References

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

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.

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, Part 2), S48–S76.

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

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

ASER (2023). Annual status of education report (aser) 2022: Provisional. Technical report, New Delhi, India.

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

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

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.

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

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

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.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

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

Cilliers, J., B. Fleisch, J. Kotze, N. Mohohlwane, S. Taylor, and T. Thulare (2022). Can virtual replace in-person coaching? Experimental evidence on teacher professional development and student learning in South Africa. *Journal of Development Economics 155*, 102815.

Cilliers, J., B. Fleisch, C. Prinsloo, and S. Taylor (2019). How to improve teaching practice? an experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*, 0618–9538R1.

Cruz, T. D. S., A. O. Loureiro, and E. Sa (2017). Full-time teachers, students, and curriculum: the single-shift model in rio de janeiro. *World Bank Policy Research Working Paper* (8086).

Dastidar, A. G. and S. Sikdar (2015). Occupation choices of high school and college students with special reference to teaching and research. *Policy Futures in Education 13*(3), 375–394.

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

Dundar, H., T. Béteille, M. Riboud, and A. Deolalikar (2014). *Student learning in South Asia: Challenges, opportunities, and policy priorities*. Washington, DC: World Bank.

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.

Ganimian, A. J., K. Muralidharan, and C. R. Walters (2023). Improving early-childhood human development: Experimental evidence from India. *Journal of Political Economy*.

Gilbert, J. B., Z. Himmelsbach, J. Soland, M. Joshi, and B. W. Domingue (2025). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *Journal of Policy Analysis and Management 44*(4), 1417–1449.

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. In T. J. Kane, K. A. Kerr, and R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. San Francisco, CA: 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 (2022a). Can education be standardized? Evidence from Kenya. (Working Paper No. 2022-68). Chicago, IL: Becker Friedman Institute For Economics.

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

Gray-Lobe, G., M. Kremer, J. de Laat, and W. Wong (2025). Details, delegation, and control: The impact of streamlining instructions to teachers on student test scores. Working paper.

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. In A. Auerbach and M. Feldstein (Eds.), *Handbook of Public Economics*, Volume 4, pp. 2045–2141. Amsterdam, the Netherlands: 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.

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

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.

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.

MHA (2012). 15th Census of India. URL: `http://www.censusindia.gov.in/` (Last accessed: February 15, 2016). Office of the Registrar General & Census Commissioner, Ministry of Home Affairs, Government of India. New Delhi, Delhi.

MoE (2024). Report on Unified District Information System for Education Plus (UDISE+) 2023-2024. New Delhi, India: Department of School Education and Literacy, Ministry of Education, Government of India. `https://www.education.gov.in/udise-report-2023-24-existing-structure`.

MoE (2026). All India survey on higher education 2023–2024. New Delhi, India: Department of Higher Education, Ministry of Education, Government of India. `https://cdnbbsr.s3waas.gov.in/s392049debbe566ca5782a3045cf300a3c/uploads/2026/07/20260708111 0623913.pdf`.

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. and V. Sundararaman (2013). Contract teachers: Experimental evidence from india. (NBER Working Paper No. 19440). Cambridge, MA: National Bureau of Economic Research (NBER).

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.

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).

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. M. Dubeck (2024). Responding to the learning crisis: Structured pedagogy in Sub-Saharan Africa. *International Journal of Educational Development 109*, 103095.

Piper, B., Y. Sitabkhan, J. Mejía, and K. Betts (2018). Effectiveness of teachers' guides in the Global South: Scripting, learning outcomes, and classroom utilization. (RTI Press Publication No. OP-0053-1805). Research Triangle Park, NC: RTI Press.

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.

Piper, B., S. S. Zuilkowski, and A. Mugenda (2014). Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. *International Journal of Educational Development 37*, 11–21.

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.

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.

Schmidt, W. H., M. T. Tatto, K. Bankov, S. Blömeke, T. Cedillo, L. Cogan, S. I. Han, R. Houang, F. J. Hsieh, L. Paine, M. Santillan, and J. Schwille (2007). *Teacher education for middle school mathematics in six countries*. East Lansing, MI: East Lansing, MI: Center for Research in Mathematics and Science Education, Michigan State University.

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: Amsterdam, the Netherlands: International Evaluation Association (IEA).

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.

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.

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.

Yoshikawa, H., D. Leyva, C. E. Snow, E. Treviño, M. C. Arbour, M. C. Barata, C. Weiland, C. Gómez, L. Moreno, A. Rolla, and N. D'Sa (2015). Experimental impacts of a teacher professional development program in chile on preschool classroom quality and child outcomes. *Journal of Developmental Psychology 51*, 309–322.

Table 1: Differences between teacher characteristics (endline)

	(1)	(2)	(3)	(4)	(5)
	Teachers		Diff. between teachers	Fellows	Diff. between teachers and fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(2)
Female	0.81	0.71	-0.10	0.63	-0.08
	[0.40]	[0.46]	(0.09)	[0.49]	(0.11)
Age	41.6	41.9	0.25	20.8	-21.0***
	[8.30]	[8.03]	(1.55)	[1.81]	(1.23)
Years of experience (total)	18.9	17.6	-1.47	1.18	-16.3***
	[8.58]	[8.29]	(1.60)	[0.49]	(1.23)
Years of experience (at this school)	5.19	5.09	-0.13	1.00	-4.09***
	[4.81]	[4.04]	(0.99)	[0.00]	(0.64)
Years of math experience (at this school)	1.51	2.53	1.01**	1.00	-1.52***
	[1.27]	[2.69]	(0.47)	[0.00]	(0.40)
Years of science experience (at this school)	1.49	2.53	1.03**	1.00	-1.52***
	[1.25]	[2.69]	(0.47)	[0.00]	(0.40)
Has bachelor's degree or higher	0.81	0.87	0.06	0.35	-0.52***
	[0.40]	[0.34]	(0.08)	[0.48]	(0.08)
Also teaches English	0.96	1.00	0.04	0.00	-1.00
	[0.20]	[0.00]	(0.03)	[0.00]	(0.00)
Also teaches Indian languages	0.89	1.00	0.11**	0.00	-1.00
	[0.31]	[0.00]	(0.05)	[0.00]	(0.00)
Teaches in English	0.06	0.02	-0.04	0.04	0.02
	[0.25]	[0.15]	(0.03)	[0.20]	(0.02)
Teaches in Hindi	0.04	0.18	0.14**	0.16	-0.02
	[0.20]	[0.39]	(0.05)	[0.37]	(0.01)
Teaches in Marathi	0.89	0.80	-0.09*	0.80	0.00
	[0.31]	[0.41]	(0.05)	[0.41]	(0.02)
Teacher assessment (IRT-scaled score)	0.00	-0.22	-0.22	1.34	1.56***
	[1.00]	[0.97]	(0.20)	[0.83]	(0.18)
Content knowledge (IRT-scaled score)	0.00	-0.06	-0.06	0.91	0.97***
	[1.00]	[1.03]	(0.23)	[0.76]	(0.20)
Instructional practices (IRT-scaled score)	0.00	-0.17	-0.17	1.12	1.29***
	[1.00]	[1.09]	(0.19)	[0.99]	(0.23)
Student errors (IRT-scaled score)	0.00	-0.25	-0.25	0.89	1.14***
	[1.00]	[0.93]	(0.18)	[1.01]	(0.18)
N (instructors)	47	45	92	49	94

Notes: This table compares teachers and fellows at endline. It shows the means and standard deviations for each group and tests for differences between groups including randomization-strata (i.e., grade) fixed effects. Column 3 compares teachers in treatment and control classrooms. Column 5 compares fellows and teachers in treatment classrooms. The sample includes all instructors 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 2: Impact on IRT-scaled all, repeated, and non-repeated items (endline, endline audit, and follow-up)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Math			Science			Language		
	All items	Repeated items	Non-repeated items	All items	Repeated items	Non-repeated items	All items	Repeated items	Non-repeated items
A. Endline									
Treatment	0.32***	0.32***	0.23***	0.23***	0.25***	0.14***	0.16***	0.15***	0.15***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Control mean		0.40	0.39		0.38	0.31		0.05	0.00
N (students)	2504	2504	2504	2497	2497	2497	2432	2432	2432
p (repeated = non-repeated)			< 0.01			< 0.01			0.84
B. Endline audit									
Treatment	0.12			0.23***					
	(0.09)			(0.06)					
N (students)	553			553					
C. Follow-up									
Treatment	0.32***	0.31***	0.22***	0.10**	0.12**	0.03	0.03	0.07**	-0.03
	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)	(0.03)
Control mean		0.32	0.30		-0.20	-0.05		-0.48	-0.36
N (students)	2270	2365	2365	2271	2367	2367	2267	2363	2363
p (repeated = non-repeated)			0.04			0.10			0.05

Notes: This table compares students' Item Response Theory (IRT)-scaled scores using all items, items in both the baseline and endline or follow-up assessments ("repeated"), and items only in the endline or follow-up assessments ("non-repeated") across the control and treatment groups. The endline and endline "audit" assessments were administered about nine months after the rollout of the intervention (April 2018), and the follow-up assessment about 11 months after the rollout of the intervention (June 2018). Repeated and non-repeated scores use only the corresponding item set. The p row tests whether the repeated- and non-repeated-item effects are equal. All specifications adjust for LASSO-selected baseline covariates. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: Impact on attendance and punctuality (unannounced school visits)

	(1)	(2)	(3)	(4)	(5)
	Teachers		Impact on teachers	Fellows	Difference between teachers and fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Attendance and punctuality					
Share of instructors who...					
...attended today	0.84	0.88	0.04	0.72	-0.12
	[0.37]	[0.33]	(0.08)	[0.45]	(0.09)
...arrived on time today	0.74	0.71	-0.03	0.44	-0.30***
	[0.44]	[0.46]	(0.09)	[0.50]	(0.08)
N (instructors)	50	49	99	50	100
B. Location at the school					
Share of present instructors who...					
...were in the classroom	0.05	0.05	0.00	0.56	0.51***
	[0.22]	[0.21]	(0.03)	[0.50]	(0.09)
...were in the principal's office	0.69	0.74	0.05	0.17	-0.52***
	[0.47]	[0.44]	(0.08)	[0.38]	(0.10)
...were elsewhere in the school	0.26	0.21	-0.05	0.28	0.02
	[0.45]	[0.41]	(0.07)	[0.45]	(0.10)
N (present instructors)	42	43	85	36	78
C. Probability of having instructor					
Share of classrooms...					
...with instructor in classroom	0.04	0.43	0.38***		
	[0.20]	[0.50]	(0.08)		
N (classrooms)	47	47	94		

Notes: This table compares average attendance and punctuality of teachers in the control and treatment groups and of 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. Panel C reports the share of classrooms in which at least one instructor was present in the classroom at the start of the school day; an instructor is the teacher in control classrooms and either the teacher or fellow in treatment classrooms. One school was not observed during the unannounced visits, so Panel C includes 47 control and 47 treatment classrooms. Standard deviations appear in brackets and standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: Impact on allocation of instructional time (announced observations)

	(1)	(2)	(3)	(4)	(5)
	Teachers		Impact on teachers	Fellows	Difference between teachers and fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time...					
...on task	0.78	0.08	-0.70***	0.75	-0.03
	[0.20]	[0.13]	(0.03)	[0.16]	(0.03)
...on class management	0.14	0.21	0.07*	0.16	0.02
	[0.13]	[0.26]	(0.04)	[0.13]	(0.02)
...off task	0.08	0.69	0.61***	0.10	0.01
	[0.14]	[0.34]	(0.05)	[0.11]	(0.02)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Minutes of lesson time...					
...on task	23.3	9.60	-13.7***	89.4	66.0***
	[5.96]	[16.1]	(2.30)	[18.7]	(2.72)
...on class management	4.16	25.0	20.8***	19.1	15.0***
	[4.02]	[31.7]	(4.29)	[15.1]	(2.16)
...off task	2.50	83.0	80.5***	11.5	9.04***
	[4.05]	[41.1]	(5.56)	[13.4]	(1.90)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of instructional time of teachers in the control and treatment groups and of fellows in the treatment group, based on announced observations about five months after the rollout of the intervention (December 2017). Observations of teachers lasted 30 minutes and those of 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 5: Impact on instructional practices (announced observations)

	(1) Teachers	(2) Fellows	(3)
	Control	Treatment	Col. (2)-(1)
A. Promoting discussion			
Allowed students to ask questions	0.39	0.58	0.20**
	[0.49]	[0.50]	(0.10)
Asked both closed and open questions	0.63	0.78	0.16*
	[0.49]	[0.42]	(0.09)
Asked students to explain their answers	0.47	0.76	0.29***
	[0.50]	[0.43]	(0.10)
Corrected student answers	0.72	0.94	0.22***
	[0.45]	[0.24]	(0.07)
Composite index (std.)	0.00	0.60	0.60***
	[1.00]	[0.71]	(0.17)
B. Fostering practice			
Assigned classwork	0.64	0.74	0.11
	[0.48]	[0.44]	(0.07)
Assigned homework	0.48	0.64	0.16*
	[0.50]	[0.49]	(0.09)
Helped individual students	0.68	0.96	0.28***
	[0.47]	[0.20]	(0.06)
Composite index (std.)	0.00	0.58	0.58***
	[1.00]	[0.77]	(0.15)
C. Cultivating climate			
Praised or encouraged students	0.76	0.96	0.20***
	[0.43]	[0.20]	(0.06)
Smiled, joked, or laughed	0.17	0.28	0.11
	[0.38]	[0.45]	(0.09)
Hit, pinched, or slapped students (reverse coded)	0.98	1.00	0.02
	[0.14]	[0.00]	(0.01)
Shouted or used harsh language (reverse coded)	0.97	0.98	0.01
	[0.18]	[0.14]	(0.03)
Composite index (std.)	0.00	0.35	0.35***
	[1.00]	[0.50]	(0.13)
D. Engaging students			
Made eye contact with nearly all students	0.75	0.74	-0.01
	[0.44]	[0.44]	(0.06)
Called on students from all seating rows	0.69	0.78	0.09
	[0.47]	[0.42]	(0.10)
Called on nearly all students by name	0.64	0.64	0.01
	[0.48]	[0.49]	(0.08)
Stayed in one spot during the lesson (reverse coded)	0.81	0.94	0.13**
	[0.39]	[0.24]	(0.06)
Composite index (std.)	0.00	0.24	0.24
	[1.00]	[0.77]	(0.16)
N (instructors)	96	50	146

Notes: This table compares the instructional practices of control teachers and fellows, based on announced observations about five months after the rollout of the intervention (December 2017). Observations of teachers lasted 30 minutes and those of fellows lasted 120 minutes. The practices are grouped by whether they promote discussion (panel A), foster practice (panel B), cultivate climate (panel C), or engage students (panel D). Variables marked as reverse coded equal one when the listed practice did not occur. The composite indexes are the first principal components of the variables in each panel, standardized with respect to control teachers. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Differences adjust for Grade fixed effects, and standard errors clustered by school appear in parentheses beneath the corresponding estimates in column (3); standard deviations appear in brackets beneath the group means in columns (1) and (2). The comparison sample includes 96 control teachers and 50 fellows observed in classrooms across the study schools during the December 2017 announced visits.

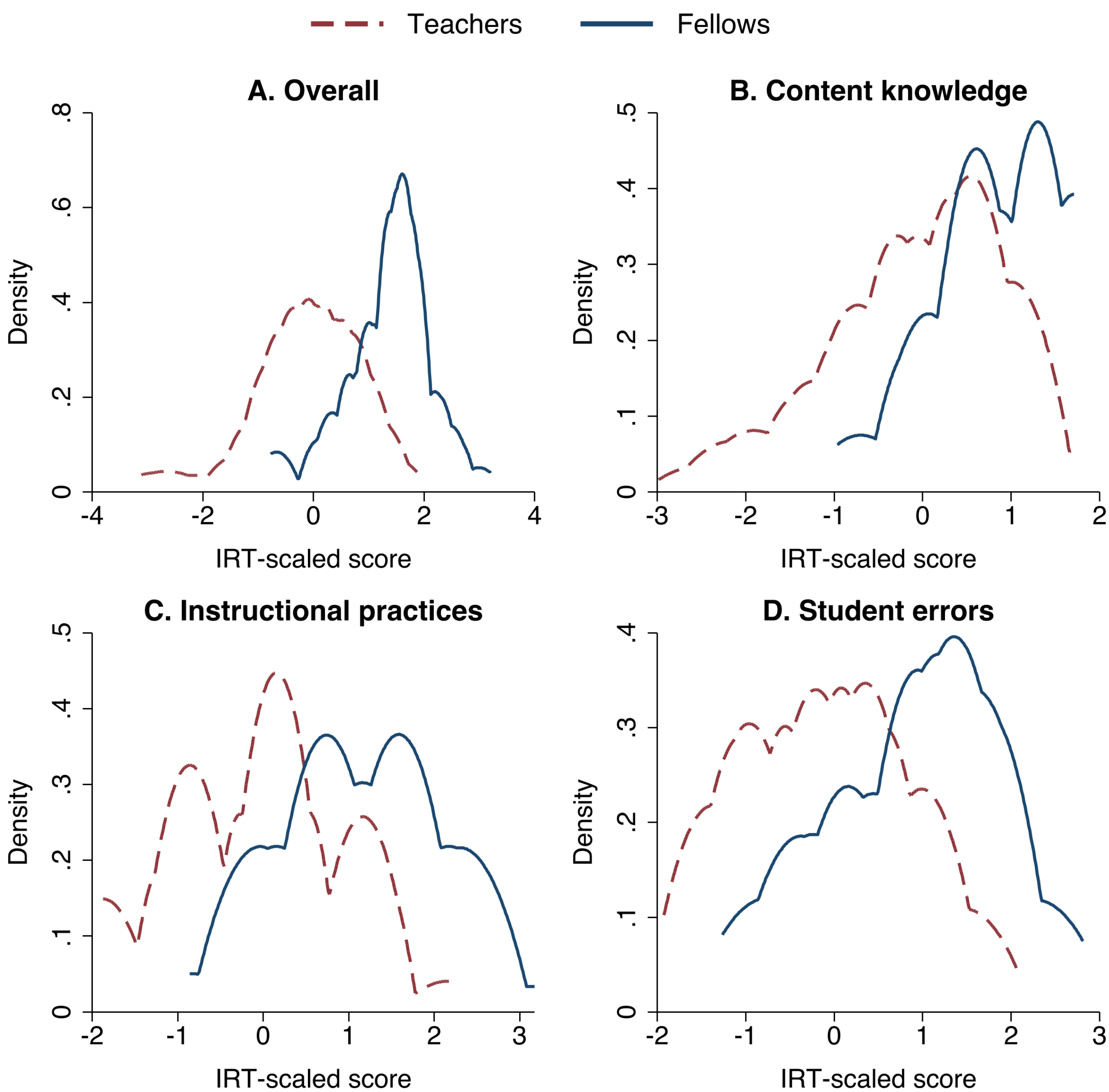
Table 6: SEI costs and implied cost-effectiveness

	(1) Estimate	(2) Unit
A. Cost inputs		
Undergraduate fellowship budget	3,662,000	INR
Fellows	60	fellows
Students served	1,920	students
Direct cost per fellow	61,033	INR
Direct cost per student	1,907	INR
Direct cost per student	30	USD
Cost per student including fixed costs	50	USD
Opportunity-cost sensitivity, per student	29.6-43.8	USD
B. Implied cost-effectiveness		
Endline math impact	0.317	SDs
Endline science impact	0.232	SDs
Math gain per USD 100, direct cost	1.06	SDs
Science gain per USD 100, direct cost	0.77	SDs
Math gain per USD 100, opportunity-cost sensitivity	0.72-1.07	SDs
Science gain per USD 100, opportunity-cost sensitivity	0.53-0.78	SDs
Math gain per USD 100, direct-cost LAYS	1.33	LAYS
Science gain per USD 100, direct-cost LAYS	0.97	LAYS
Math gain per USD 100, opportunity-cost LAYS	0.90-1.34	LAYS
Science gain per USD 100, opportunity-cost LAYS	0.66-0.98	LAYS
C. Smart Buys benchmarks		
Targeted instruction or structured lesson plans, average	3.0	LAYS per USD 100
Targeted instruction in India, upper benchmark	4.0	LAYS per USD 100

Notes: This table reports cost inputs and implied cost-effectiveness estimates for the Science Education Initiative (SEI). Panel A reports costs from SEI budget records for the 2017-2018 undergraduate fellowship program, except the fixed-cost and opportunity-cost rows. The number of fellows and students is implied by the total undergraduate fellowship budget, the reported cost per fellow, and the reported cost per student. The fixed-cost row comes from an earlier SEI working paper. The opportunity-cost range comes from a worksheet that adds fellows' opportunity cost under alternative wage assumptions. Panel B divides the covariate-adjusted endline impacts in Table 2 by the cost per student and multiplies by 100. The direct-cost rows use INR 1,907 per student, rounded to USD 30 in the source materials; the opportunity-cost rows use the USD 29.6-43.8 range from the opportunity-cost worksheet. LAYS rows convert SD gains to learning-adjusted years of schooling by dividing by the 0.8 SD high-quality benchmark learning rate used by Angrist et al. (2020). Panel C reports comparison benchmarks from Angrist et al. (2020), which informed the Smart Buys classifications in Akyeampong et al. (2023). These benchmarks are included for order-of-magnitude comparison rather than precise ranking.

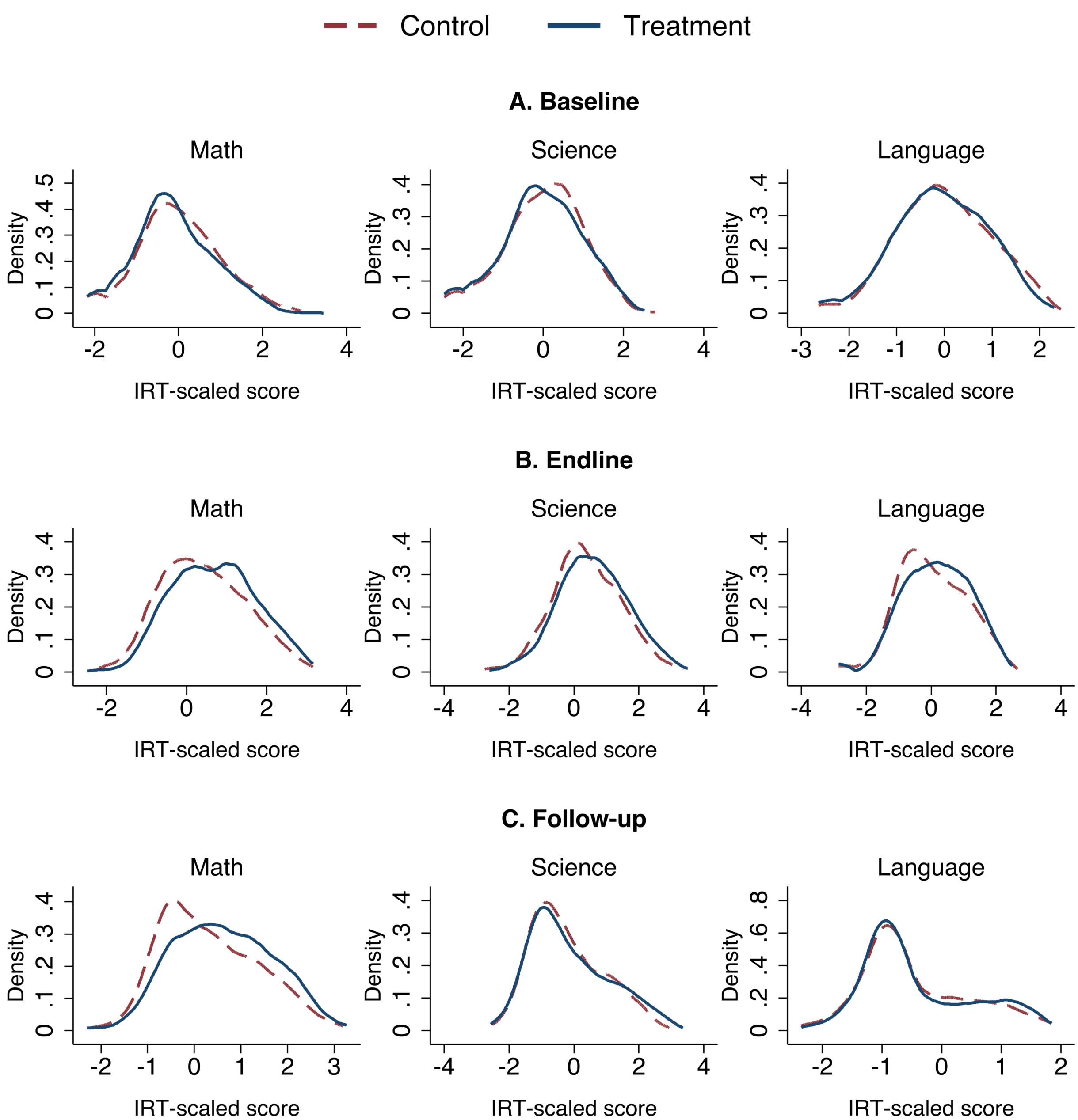
Appendix A Additional tables and figures

Figure A.1: Distributions of IRT-scaled scores on teacher assessments



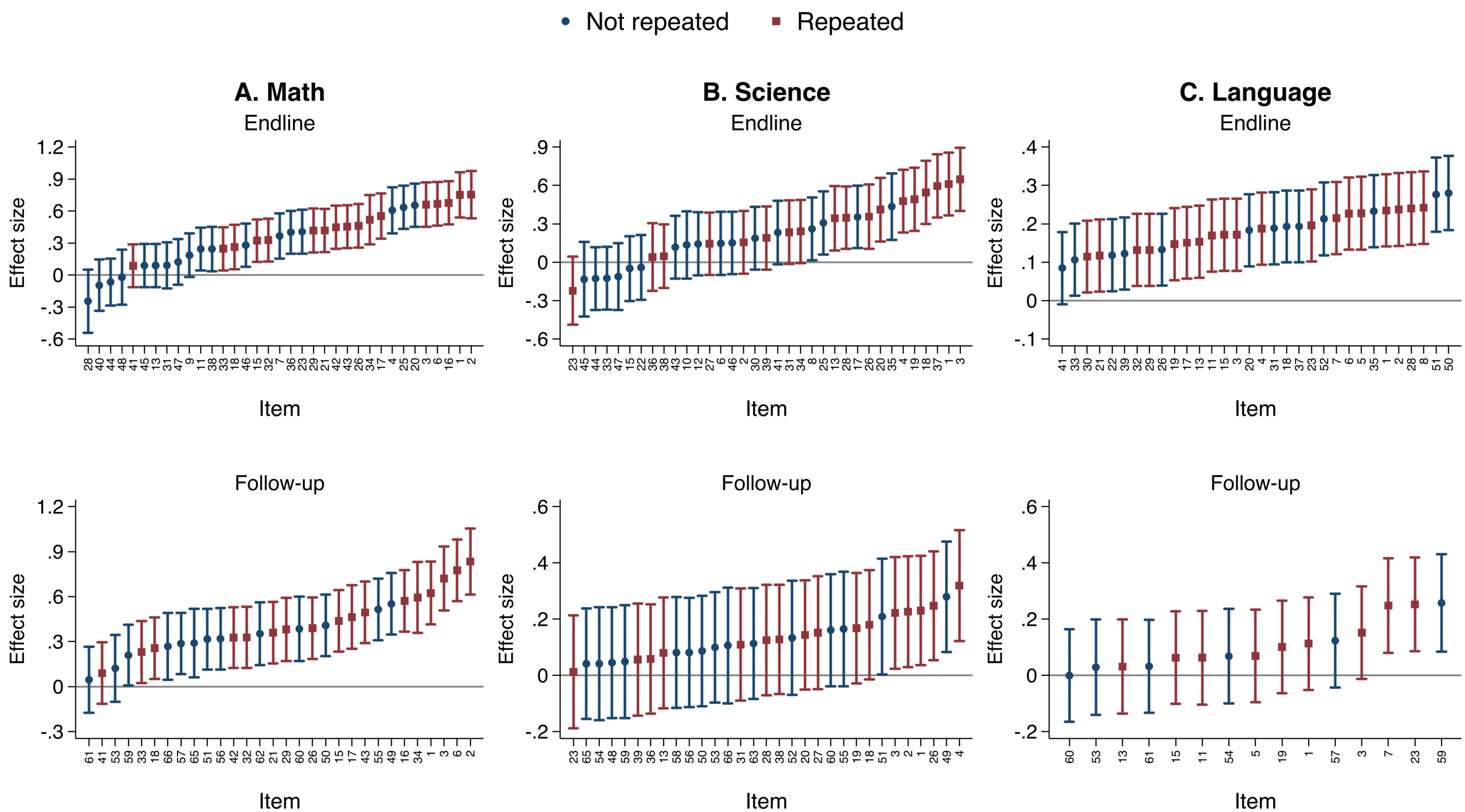
Notes: The figure shows the distributions of the overall and domain-specific IRT-scaled scores on the teacher assessment. Scores are estimated using Rasch models and standardized to have mean zero and standard deviation one among control-group teachers, with higher values indicating better performance.

Figure A.2: Distributions of IRT-scaled scores on student assessments



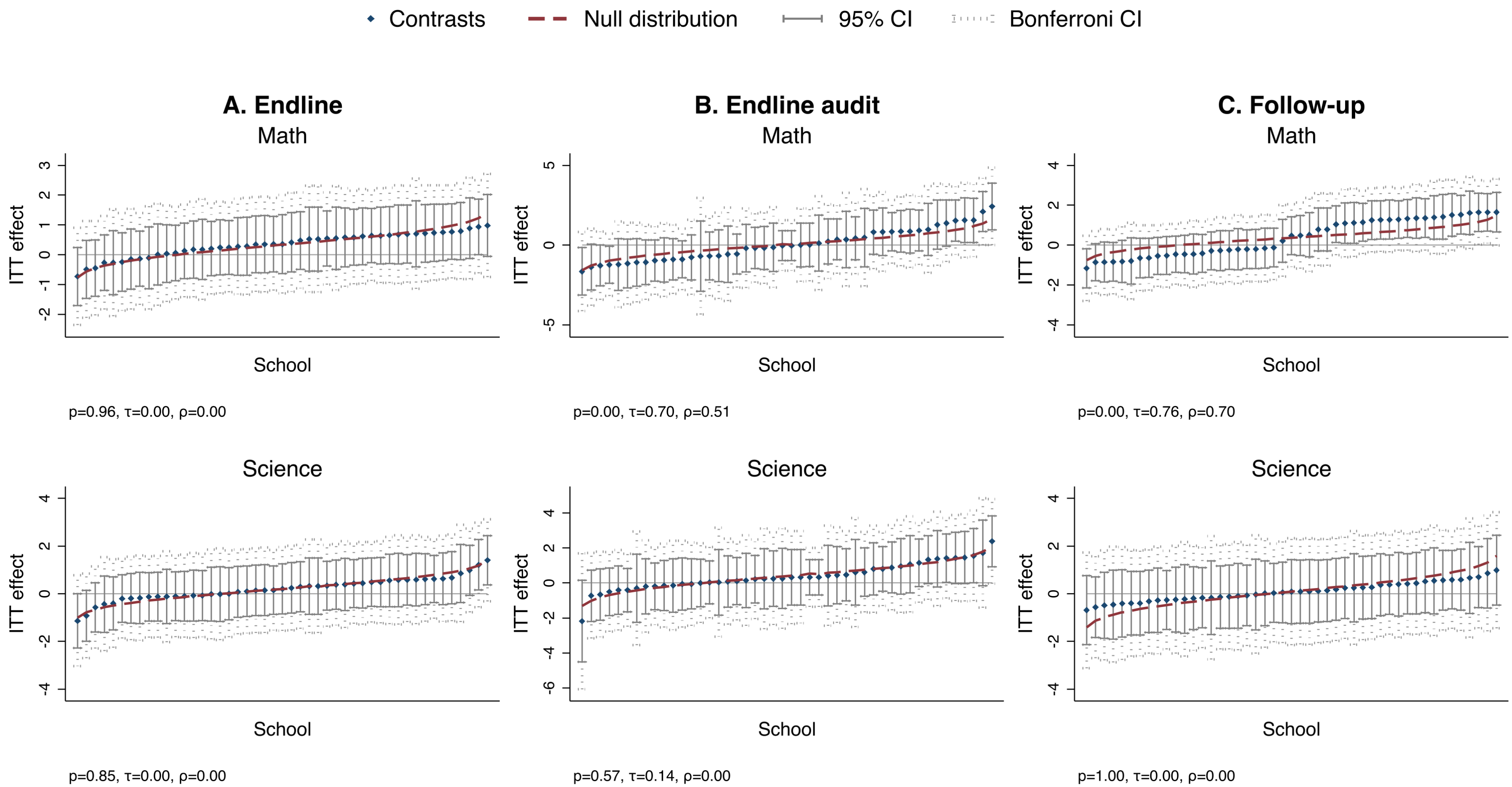
Notes: The figure shows the distributions of linked IRT-scaled scores on the student assessments. Scores are estimated using two-parameter logistic IRT models, with higher values indicating better performance.

Figure A.3: Item-specific treatment effects



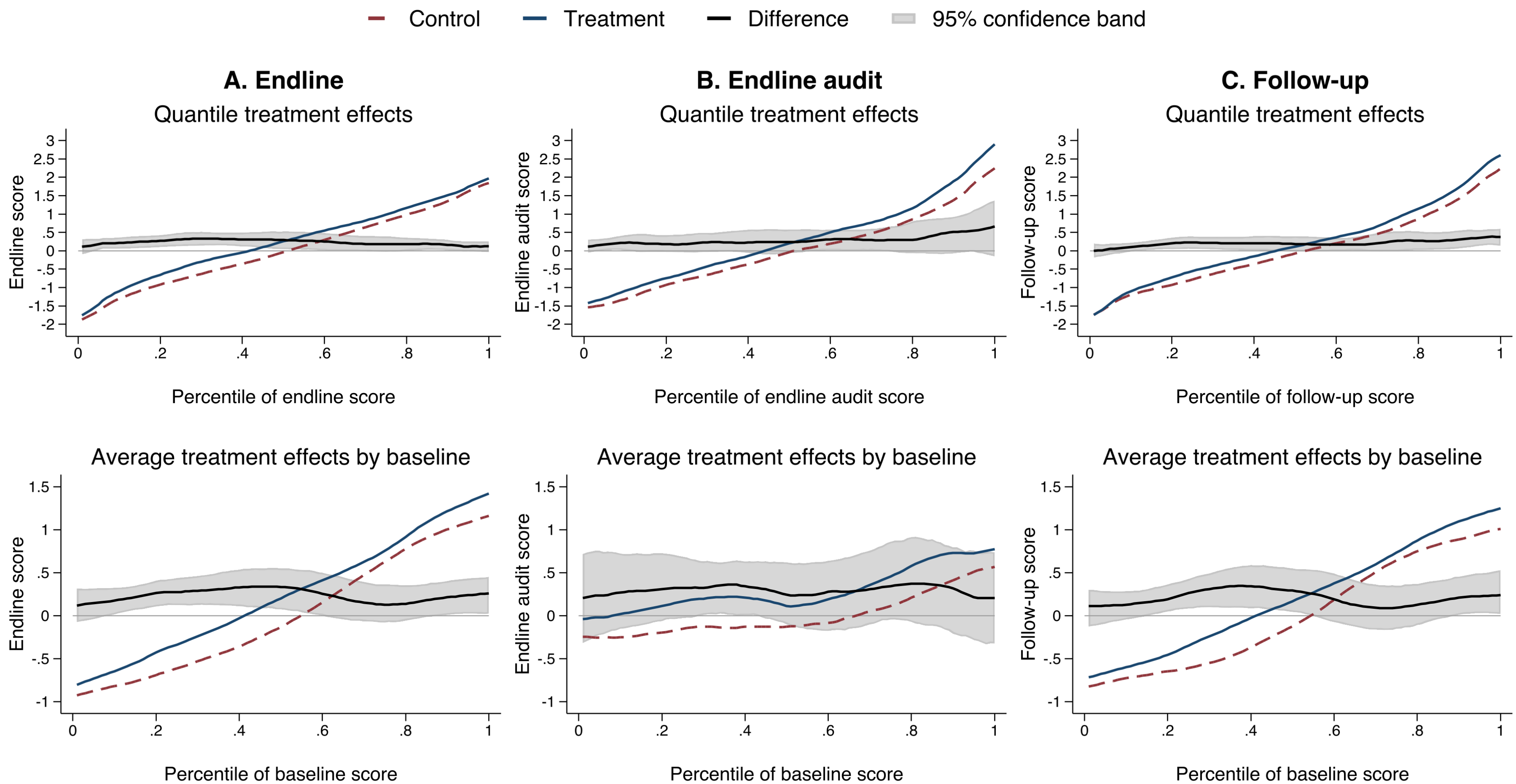
Notes: This figure shows empirical Bayes estimates of item-specific treatment effects on the endline and follow-up assessments in math, science, and language, with 95% confidence intervals. Maroon squares indicate items repeated from baseline; navy circles indicate items introduced after baseline. The estimates are regularized effects for the items in each assessment.

Figure A.4: 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 standardized scores in math and science by school at endline and the 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). Each navy diamond refers to the point estimate for a given school. Bonferroni confidence intervals adjust standard errors for multiple hypothesis testing. The maroon dashed line shows the null distribution of "effects" that can be expected due to error. The p-values test whether school-specific treatment effects are homogeneous within each subject-round combination displayed across the six panels in the figure.

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



Notes: The top row shows quantiles of endline, endline audit, and follow-up composite scores for treatment and control students who participated in the baseline and each corresponding round, estimated by polynomial regressions of each round's scores on that round's percentiles separately by experimental group. The bottom row shows average scores and treatment effects at each percentile of the baseline composite score for the same samples, estimated by polynomial regression. Composite scores are the first principal component from an analysis of all subjects assessed in each round. Gray bands show bootstrapped 95% confidence intervals; solid black lines show the estimated treatment effects.

Table A.1: Differences between student characteristics (baseline)

	(1) Control	(2) Treatment	(3) Col. (2)-(1)
Female	0.51	0.54	0.04
	[0.50]	[0.50]	(0.03)
Age	10.8	10.9	0.06
	[1.01]	[1.04]	(0.05)
Speaks Marathi at home	0.69	0.69	0.00
	[0.46]	[0.46]	(0.02)
Speaks Hindi at home	0.17	0.16	-0.01
	[0.38]	[0.37]	(0.02)
Speaks English at home	0.01	0.01	0.00
	[0.10]	[0.08]	(0.01)
Mother completed primary school	0.63	0.60	-0.03
	[0.48]	[0.49]	(0.02)
Father completed primary school	0.77	0.78	0.01
	[0.42]	[0.42]	(0.02)
Student has a desk to study	0.28	0.26	-0.02
	[0.45]	[0.44]	(0.03)
Student has own room	0.17	0.20	0.03
	[0.38]	[0.40]	(0.03)
Student has a computer	0.20	0.18	-0.02
	[0.40]	[0.38]	(0.03)
Student has Internet	0.43	0.37	-0.05
	[0.50]	[0.48]	(0.05)
Student has a TV	0.90	0.89	-0.01
	[0.30]	[0.31]	(0.02)
Attends tuition in math	0.35	0.32	-0.03
	[0.48]	[0.47]	(0.04)
Attends tuition in science	0.26	0.23	-0.03
	[0.44]	[0.42]	(0.04)
Attends tuition in Marathi or Hindi	0.37	0.36	-0.01
	[0.48]	[0.48]	(0.03)
Attends tuition in English	0.33	0.31	-0.02
	[0.47]	[0.46]	(0.03)
Math (standardized score)	0.00	-0.13	-0.13*
	[1.00]	[0.96]	(0.07)
Science (standardized score)	0.00	-0.05	-0.05
	[1.00]	[1.01]	(0.07)
Language (standardized score)	0.00	-0.07	-0.07
	[1.00]	[1.02]	(0.06)
Raven's matrices (standardized score)	0.00	0.05	0.05
	[1.00]	[0.98]	(0.06)
N (students)	1,367	1,290	2,657

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. The p-value from a joint F-test of all listed covariates in a regression of treatment status on those covariates, with randomization-strata fixed effects and standard errors clustered by school, is 0.013 (0.159 when math achievement is excluded). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.2: Attrition rates in assessments (endline and follow-up)

	(1) Endline	(2) Follow-up
A. Treatment		
Treatment	-0.02	0.00
	(0.01)	(0.02)
N (students)	2657	2657
Control mean	0.14	0.24
B. Treatment and baseline		
Treatment	0.16	0.18
	(0.17)	(0.20)
Female	-0.08**	-0.06*
	(0.04)	(0.04)
Age	0.03**	0.04**
	(0.01)	(0.02)
Composite score (at baseline)	-0.05***	-0.04***
	(0.01)	(0.01)
Female × Treatment	0.06	0.05
	(0.04)	(0.04)
Age × Treatment	-0.02	-0.02
	(0.02)	(0.02)
Composite score × Treatment	-0.01	-0.01
	(0.01)	(0.01)
N (students)	2404	2404
F-ratio (Interactions)	1.69	1.10
p (interactions = 0)	0.18	0.36

Notes: This table shows estimates from regressions predicting whether students have observed test scores in assessments administered at endline, 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). The outcome is defined as having an observed test score in the corresponding assessment round. The sample includes students present at baseline. Panel A regresses this indicator on treatment status, and Panel B regresses this indicator on treatment status interacted with baseline characteristics. The p row tests whether all interaction terms equal zero. Both panels include randomization-strata fixed effects. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Script adherence among fellows (announced observations)

	(1) All fellows	(2) High-scoring	(3) Low-scoring
Share of fellows who...			
...wrote on blackboard as indicated in the script	0.84	0.83	0.84
	[0.37]	[0.38]	[0.38]
...asked students questions in the script	0.79	0.87	0.68
	[0.42]	[0.34]	[0.48]
...used materials as indicated in the script	0.58	0.55	0.61
	[0.50]	[0.51]	[0.50]
...assigned students activities in the script	0.52	0.65	0.37
	[0.51]	[0.49]	[0.50]
...changed/rephrased parts of the script	0.34	0.28	0.42
	[0.48]	[0.46]	[0.51]
...added to or expanded on parts of the script	0.34	0.24	0.47
	[0.48]	[0.44]	[0.51]
...excluded parts of the script	0.18	0.16	0.21
	[0.39]	[0.37]	[0.42]
N (fellows)	50	28	22

Notes: This table displays the prevalence of script adherence among 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 at or above the fellow-specific median of the overall IRT-scaled teacher-assessment score (described in section 3); low-scoring fellows are those below that median.

Table A.4: SEI-specific fellow survey responses (endline)

	(1) Share	(2) SD	N (fellows)
A. Views on SEI fellowship			
Teaching method was very good	0.92	0.28	49
Program followed the curriculum	0.78	0.42	49
Program provided quality instruction	0.96	0.20	49
Materials caused no problems	0.96	0.20	49
Training was sufficient	0.96	0.20	49
Program improved science and math learning	0.98	0.14	49
B. Self-efficacy			
Knew material and was well prepared	0.80	0.41	49
Was prepared to manage a classroom	0.76	0.43	49
Felt prepared to teach in the classroom	0.96	0.20	49
C. Challenges			
Reported challenges with teachers	0.27	0.45	49
Reported challenges with school leadership	0.12	0.33	49
Reported that school was too far	0.37	0.49	49
Reported that students did not respect them	0.20	0.41	49
Reported financial constraints	0.18	0.39	49
D. Career			
Reported they would choose fellowship again	0.92	0.28	49
Reported interest in a teaching career	0.82	0.39	49
Reported expecting a full-time job offer	0.84	0.37	49

Notes: This table reports endline responses by fellows to SEI-specific survey questions. The sample includes fellows who responded to the endline teacher and fellow survey. Column 1 reports shares of fellows. Panels A and B report the share who agree or strongly agree with each statement; negative statements are reverse-coded, so higher values indicate more favorable views of SEI or greater self-efficacy. The material and training rows in Panel A are reverse-coded challenge items and report the share not reporting each challenge. Panel C reports the share reporting each challenge. Panel D reports the share who would choose the SEI fellowship again, are interested or very interested in a teaching career, and expect a somewhat or very likely full-time job offer from SEI. Standard deviations appear in column 2.

Table A.5: Round of data collection for the study

(1)	(2)	(3)	(4)	(5)	(6)
			Student participation rates		
Month	Event	Grades	Both groups	Control	Treatment
A. 2017-2018 school year					
June	School year starts				
Baseline:					
July	Student assessments	5, 6	100%	100%	100%
	Student surveys	5, 6	100%	100%	100%
Midline:					
December	Announced classroom observations	5, 6	99%	98%	99%
February	Unannounced school visits	5, 6	99%	98%	99%
	Administrative data on student attendance	5, 6	99%	98%	99%
Endline:					
March-	Student assessments	5, 6			
April	− Math and science		87%	86%	88%
	− Language		87%	86%	88%
	Student surveys	5, 6	78%	77%	80%
	Student "audit" assessments of math and science	5, 6	21%	21%	21%
April-	Teacher surveys	5, 6	95%	94%	95%
May	Teacher assessments	5, 6	93%	92%	94%
B. 2018-2019 school year					
June	School year starts				
Follow-up:					
June-	Student assessments	6, 7			
July	− Math		76%	76%	76%
	− Science		76%	76%	76%
	− Language		76%	76%	76%
August	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 A.6: Impact on proportion-correct test scores (endline, endline audit, and follow-up)

	(1) Math	(2) Science	(3) Language
A. Endline			
Treatment	0.06***	0.04***	0.04***
	(0.01)	(0.01)	(0.01)
N (students)	2524	2524	2524
B. Endline audit			
Treatment	0.02	0.10***	
	(0.03)	(0.03)	
N (students)	553	553	
C. Follow-up			
Treatment	0.06***	0.02***	0.01*
	(0.01)	(0.01)	(0.00)
N (students)	2369	2369	2369

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 expressed as proportions of items answered correctly. Impact estimates include randomization-strata fixed effects and LASSO-selected baseline covariates. 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 by topic and skill

	(1)	(2) Topics	(3)	(4)	(5) Skills	(6)
	Numbers	Geometry and measurement	Data display	Knowing	Applying	Reasoning
A. Math						
Endline						
Treatment	0.27***	0.26***	0.11**	0.25***	0.30***	0.06
	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.05)
N (students)	2504	2504	2504	2504	2504	2504
Follow-up						
Treatment	0.32***	0.23***	0.16***	0.31***	0.26***	0.16***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)
N (students)	2270	2270	2270	2270	2270	2270
B. Science						
	Life science	Earth science	Physical science	Knowing	Applying	Reasoning
Endline						
Treatment	0.21***	0.19***	0.13***	0.20***	0.18***	0.14***
	(0.05)	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)
N (students)	2497	2497	2497	2497	2497	2497
Follow-up						
Treatment	0.08*	0.11**	0.12**	0.10**	0.09*	0.14***
	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)	(0.05)
N (students)	2271	2271	2271	2271	2271	2271
C. Language						
	Vocabulary	Grammar	Reading	Retrieving information	Making inferences	Interpreting and integrating
Endline						
Treatment	0.13***	0.10**	0.16***	0.13***	0.16***	0.08
	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.05)
N (students)	2432	2432	2432	2432	2432	2432
Follow-up						
Treatment	0.10***	0.01	0.00	0.08**	0.02	0.03
	(0.04)	(0.04)	(0.05)	(0.03)	(0.04)	(0.04)
N (students)	2267	2267	2267	2267	2267	2267

Notes: This table compares students' topic- and skill-specific test scores in the control and treatment groups based on assessments administered at endline, about nine months after the rollout of the intervention (April 2018), and at follow-up, about 11 months after the rollout (June 2018). Scores come from one-parameter logistic Item Response Theory models estimated separately by topic and skill using pooled item responses from baseline, endline, and follow-up; repeated items link scores across rounds. Scores are standardized to have mean zero and standard deviation one in the control group within each round. Impact estimates come from regressions that include grade fixed effects and LASSO-selected baseline covariates. Standard errors clustered by school appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.8: Heterogeneous impact on IRT-scaled test scores by student characteristics (endline, endline audit, and follow-up)

	(1) Baseline score			(2) SES index			(3) Female			(4) Scheduled caste/tribe		
	Math	Science	Language	Math	Science	Language	Math	Science	Language	Math	Science	Language
A. Endline												
Treatment	0.32***	0.19***	0.14***	0.34***	0.23***	0.15***	0.28***	0.18***	0.13**	0.33***	0.18***	0.10
	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.06)	(0.05)	(0.05)	(0.06)	(0.07)
Covariate	0.54***	0.41***	0.57***	0.01	0.00	-0.03	0.01	0.07	0.16***	-0.10**	-0.12**	-0.09
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.07)	(0.05)	(0.05)	(0.05)	(0.06)
Treatment × covariate	0.06	0.04	-0.04	0.05	0.09**	0.06**	0.08	0.10	0.06	0.02	0.11	0.12
	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.08)	(0.11)	(0.09)	(0.06)	(0.07)	(0.09)
p (main + interaction = 0)	< 0.01	< 0.01	< 0.01	0.06	0.05	0.36	0.19	0.03	< 0.01	0.10	0.75	0.72
N (students)	2,504	2,497	2,432	2,236	2,232	2,183	2,504	2,497	2,432	2,217	2,215	2,163
B. Endline audit												
Treatment	0.13	0.23***		0.13	0.26***		0.07	0.33***		0.24*	0.34***	
	(0.09)	(0.06)		(0.09)	(0.06)		(0.10)	(0.06)		(0.13)	(0.07)	
Covariate	0.11*	0.07		0.03	-0.04		0.06	0.18**		0.02	0.07	
	(0.06)	(0.04)		(0.09)	(0.03)		(0.13)	(0.08)		(0.14)	(0.06)	
Treatment × covariate	0.04	0.04		-0.06	0.04		0.10	-0.15		-0.22	-0.14*	
	(0.09)	(0.05)		(0.08)	(0.05)		(0.16)	(0.10)		(0.16)	(0.07)	
p (main + interaction = 0)	0.03	0.01		0.67	0.96		0.24	0.71		0.06	0.34	
N (students)	553	553		532	532		553	553		537	537	
C. Follow-up												
Treatment	0.33***	0.08*	0.02	0.35***	0.13**	0.05	0.30***	0.09	0.04	0.29***	0.12	0.04
	(0.05)	(0.05)	(0.03)	(0.04)	(0.05)	(0.04)	(0.06)	(0.08)	(0.06)	(0.05)	(0.08)	(0.05)
Covariate	0.54***	0.21***	0.30***	0.00	-0.06	-0.02	0.06	0.11	0.07	-0.11**	-0.07	-0.05
	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.09)	(0.07)	(0.05)	(0.08)	(0.06)
Treatment × covariate	0.03	0.05	0.02	0.03	0.08*	0.04	0.05	0.02	-0.03	0.10	-0.02	0.02
	(0.05)	(0.05)	(0.04)	(0.03)	(0.04)	(0.03)	(0.07)	(0.11)	(0.09)	(0.07)	(0.10)	(0.08)
p (main + interaction = 0)	< 0.01	< 0.01	< 0.01	0.43	0.50	0.51	0.06	0.10	0.47	0.80	0.21	0.53
N (students)	2,270	2,271	2,267	1,880	1,881	1,879	2,270	2,271	2,267	1,863	1,864	1,862

Notes: This table shows the impact of the fellowship on students' IRT-scaled math, science, and language test scores by baseline achievement, socio-economic status, sex, and caste. Panels A-C report results from the endline, endline audit, and follow-up assessments, respectively. IRT-scaled scores are standardized to have mean zero and standard deviation one in the control group. The p row tests whether the main fellowship effect plus its interaction with the indicated covariate equals zero. All specifications include randomization-strata fixed effects and LASSO-selected baseline covariates. Standard errors clustered by school appear in parentheses. Significance levels: * 10%; ** 5%; *** 1%.

Table A.9: Heterogeneous impact on IRT-scaled test scores by instructor characteristics (endline, endline audit, and follow-up)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Sex match			Instructor score			Fellow-teacher score gap		
	Math	Science	Language	Math	Science	Language	Math	Science	Language
A. Endline									
Treatment	0.36***	0.25***	0.16**	0.42***	0.27***	0.20***	0.36***	0.19**	0.14**
	(0.06)	(0.06)	(0.06)	(0.06)	(0.08)	(0.07)	(0.06)	(0.08)	(0.06)
Covariate	-0.02	0.08	0.07	0.06	0.08*	-0.04	-0.01	-0.03	-0.02
	(0.05)	(0.07)	(0.06)	(0.04)	(0.04)	(0.08)	(0.04)	(0.04)	(0.03)
Treatment × covariate	-0.01	-0.01	0.00	-0.12**	-0.10	0.01	-0.02	0.03	-0.01
	(0.08)	(0.10)	(0.09)	(0.05)	(0.06)	(0.08)	(0.04)	(0.05)	(0.03)
p (main + interaction = 0)	0.64	0.35	0.33	0.08	0.58	0.59	< 0.01	< 0.01	< 0.01
N (students)	2,113	2,109	2,060	2,358	2,351	2,286	2,242	2,235	2,167
B. Endline audit									
Treatment	0.07	0.31***		0.14	0.21**		0.17	0.41***	
	(0.12)	(0.09)		(0.18)	(0.09)		(0.18)	(0.10)	
Covariate	-0.04	0.16**		-0.12	-0.05		-0.04	0.07*	
	(0.14)	(0.07)		(0.08)	(0.04)		(0.07)	(0.04)	
Treatment × covariate	0.08	-0.07		0.06	0.07		-0.07	-0.13*	
	(0.17)	(0.12)		(0.15)	(0.08)		(0.09)	(0.06)	
p (main + interaction = 0)	0.72	0.31		0.60	0.75		0.34	< 0.01	
N (students)	506	506		523	523		499	499	
C. Follow-up									
Treatment	0.41***	0.10	0.07	0.42***	0.07	0.03	0.35***	0.14**	0.01
	(0.07)	(0.07)	(0.06)	(0.09)	(0.07)	(0.05)	(0.07)	(0.07)	(0.05)
Covariate	0.04	0.06	0.04	0.08	-0.01	0.01	-0.03	0.03	0.02
	(0.06)	(0.07)	(0.07)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)
Treatment × covariate	0.00	0.05	0.00	-0.09	0.03	0.00	-0.01	-0.05	0.00
	(0.10)	(0.09)	(0.08)	(0.08)	(0.06)	(0.04)	(0.04)	(0.04)	(0.02)
p (main + interaction = 0)	0.57	0.12	0.52	0.78	0.61	0.67	< 0.01	0.08	0.73
N (students)	1,781	1,782	1,780	2,142	2,143	2,139	2,037	2,038	2,034

Notes: This table shows the impact of the fellowship on students' IRT-scaled math, science, and language test scores by student-instructor sex match, instructor assessment score, and fellow-teacher assessment-score gap. Panels A-C report endline, endline audit, and follow-up results, respectively. The score gap is the fellow's score minus the teacher's score, collapsed to the school-shift level; its interaction should be interpreted descriptively. Scores are standardized to have mean zero and standard deviation one in their respective control groups. The p row tests whether the main fellowship effect plus its interaction equals zero. All specifications include randomization-strata fixed effects and LASSO-selected baseline covariates. Standard errors clustered by school appear in parentheses. Significance levels: * 10%; ** 5%; *** 1%.

Table A.10: Impact on student attitudes, mindsets, and aspirations (endline)

	(1) Control	(2) Treatment	(3) Difference	N (students)
A. Math and science				
Enjoy learning math	0.93	0.93	0.00	2,231
	[0.26]	[0.25]	(0.01)	
Enjoy learning science	0.85	0.87	0.02	2,225
	[0.36]	[0.34]	(0.03)	
Wish they did not have math	0.56	0.62	0.06*	2,211
	[0.50]	[0.49]	(0.03)	
Wish they did not have science	0.56	0.62	0.05*	2,204
	[0.50]	[0.49]	(0.03)	
Feel nervous about math	0.40	0.41	0.01	2,197
	[0.49]	[0.49]	(0.03)	
Feel nervous about science	0.39	0.39	0.00	2,197
	[0.49]	[0.49]	(0.03)	
Find math useful for life	0.83	0.86	0.02	2,223
	[0.37]	[0.35]	(0.02)	
Find science useful for life	0.83	0.81	-0.01	2,228
	[0.38]	[0.39]	(0.02)	
Stop trying when math gets hard	0.67	0.64	-0.04	2,221
	[0.47]	[0.48]	(0.03)	
Stop trying when science gets hard	0.67	0.66	-0.02	2,218
	[0.47]	[0.48]	(0.03)	
Composite index (std.)	0.00	0.08	0.07	2,016
	[1.00]	[0.98]	(0.07)	
B. Intelligence				
People cannot change their intelligence	0.54	0.52	-0.02	2,524
	[0.50]	[0.50]	(0.03)	
People have a fixed amount of intelligence	0.46	0.45	-0.01	2,524
	[0.50]	[0.50]	(0.03)	
Only some people are people	0.44	0.42	-0.01	2,524
	[0.50]	[0.49]	(0.03)	
Boys are more intelligent than girls	0.39	0.37	-0.02	2,524
	[0.49]	[0.48]	(0.03)	
Boys are better at math and science	0.37	0.36	-0.01	2,524
	[0.48]	[0.48]	(0.03)	
Composite index (std.)	0.00	-0.05	-0.05	2,524
	[1.00]	[1.02]	(0.06)	
C. Aspirations				
Want to continue studying after high school	0.68	0.70	0.02	2,246
	[0.47]	[0.46]	(0.02)	
Want to study a STEM subject in high school	0.76	0.80	0.04	2,239
	[0.43]	[0.40]	(0.03)	
Want a STEM-related job	0.33	0.33	0.00	2,231
	[0.47]	[0.47]	(0.03)	
Composite index (std.)	0.00	0.04	0.04	2,216
	[1.00]	[0.98]	(0.05)	

Notes: This table compares students' attitudes, mindsets, 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.11: Heterogeneous impact on student beliefs by student and instructor characteristics (endline)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Female			Baseline score			SES index			Scheduled caste/tribe		
	STEM	Intelligence	Aspirations	STEM	Intelligence	Aspirations	STEM	Intelligence	Aspirations	STEM	Intelligence	Aspirations
A. Student characteristics												
Treatment	0.07	-0.08	0.14*	0.10	-0.10	0.09	0.09	-0.09	0.10*	0.07	-0.17**	0.14**
	(0.10)	(0.08)	(0.08)	(0.07)	(0.07)	(0.05)	(0.07)	(0.07)	(0.05)	(0.09)	(0.07)	(0.07)
Covariate	0.08	-0.47***	0.21**	0.18***	-0.08***	0.09***	-0.09	0.10**	0.03	-0.02	-0.07	0.05
	(0.09)	(0.09)	(0.09)	(0.03)	(0.02)	(0.02)	(0.06)	(0.04)	(0.04)	(0.05)	(0.05)	(0.07)
Treatment × covariate	0.03	-0.02	-0.07	0.00	0.02	0.01	0.10*	-0.12**	0.00	0.04	0.11	-0.08
	(0.12)	(0.13)	(0.10)	(0.03)	(0.04)	(0.03)	(0.05)	(0.05)	(0.05)	(0.08)	(0.09)	(0.09)
p (main + interaction = 0)	0.21	< 0.01	0.03	< 0.01	0.08	< 0.01	0.81	0.64	0.51	0.73	0.56	0.62
N (students)	1,825	1,973	2,009	1,862	2,015	2,052	1,824	1,970	2,006	1,804	1,953	1,989
	Sex match			Instructor score			Fellow-teacher score gap					
	STEM	Intelligence	Aspirations	STEM	Intelligence	Aspirations	STEM	Intelligence	Aspirations			
B. Instructor characteristics												
Treatment	0.17*	-0.19*	0.09	0.08	-0.12	0.07	0.03	-0.06	0.15**			
	(0.10)	(0.10)	(0.09)	(0.09)	(0.11)	(0.07)	(0.11)	(0.11)	(0.07)			
Covariate	0.09	-0.34***	0.05	0.06	-0.05	-0.02	0.02	0.02	0.02			
	(0.08)	(0.11)	(0.09)	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)			
Treatment × covariate	-0.10	0.18	0.02	-0.03	0.07	0.01	0.04	-0.05	-0.09**			
	(0.14)	(0.14)	(0.11)	(0.06)	(0.07)	(0.07)	(0.07)	(0.06)	(0.04)			
p (main + interaction = 0)	0.88	0.15	0.18	0.45	0.78	0.71	0.17	0.58	0.08			
N (students)	1,722	1,864	1,895	1,899	2,057	2,086	1,806	1,955	1,983			

Notes: This table shows the impact of the fellowship on students' STEM beliefs, intelligence beliefs, and aspirations by student and instructor characteristics. Panel A reports heterogeneity by student sex, baseline achievement, socio-economic status, and caste. Panel B reports heterogeneity by student-instructor sex match, instructor assessment score, and the fellow-teacher assessment-score gap. The outcomes were measured in surveys administered at endline. Indexes are standardized to have mean zero and standard deviation one in the control group. The fellow-teacher score gap is collapsed to the school-shift level, so its interaction should be interpreted descriptively. The p row tests whether the main fellowship effect plus its interaction with the indicated covariate equals zero. All specifications include randomization-strata fixed effects and LASSO-selected baseline covariates. Standard errors clustered by school appear in parentheses. Significance levels: * 10%; ** 5%; *** 1%.

Table A.12: Impact on student attendance (unannounced visits, school registers, and endline)

	(1) Control	(2) Treatment	(3) Difference	N (students)
A. Unannounced visits				
Share of students observed at school	0.79	0.79	0.00	2,442
	[0.41]	[0.41]	(0.03)	
B. School registers				
Absent 0 days this week	0.55	0.54	-0.01	2,619
	[0.50]	[0.50]	(0.03)	
Absent 1-2 days this week	0.27	0.27	0.00	2,619
	[0.45]	[0.45]	(0.02)	
Absent 3-4 days this week	0.07	0.08	0.01	2,619
	[0.26]	[0.28]	(0.01)	
Absent 5+ days this week	0.10	0.10	0.00	2,619
	[0.30]	[0.30]	(0.02)	
C. Endline student survey				
Tardiness				
Late 0 days this week	0.40	0.39	-0.01	2,219
	[0.49]	[0.49]	(0.03)	
Late 1-2 days this week	0.30	0.34	0.04**	2,219
	[0.46]	[0.47]	(0.02)	
Late 3-4 days this week	0.17	0.14	-0.03	2,219
	[0.37]	[0.35]	(0.02)	
Late 5+ days this week	0.14	0.14	0.00	2,219
	[0.34]	[0.34]	(0.02)	
Attendance				
Absent 0 days this week	0.41	0.41	0.00	2,222
	[0.49]	[0.49]	(0.03)	
Absent 1-2 days this week	0.33	0.31	-0.03	2,222
	[0.47]	[0.46]	(0.02)	
Absent 3-4 days this week	0.13	0.15	0.02	2,222
	[0.33]	[0.36]	(0.02)	
Absent 5+ days this week	0.13	0.13	0.00	2,222
	[0.34]	[0.34]	(0.02)	

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.13: Impact on student demand for tuition (endline)

	(1) Control	(2) Treatment	(3) Difference	N (students)
A. Subjects				
In math	0.32	0.35	0.03	2,195
	[0.47]	[0.48]	(0.03)	
In science	0.26	0.29	0.03	2,173
	[0.44]	[0.45]	(0.03)	
In language	0.24	0.27	0.03	2,151
	[0.43]	[0.44]	(0.03)	
In English	0.32	0.35	0.03	2,196
	[0.47]	[0.48]	(0.03)	
B. Duration				
Less than 2 hours per week	0.04	0.04	0.00	2,237
	[0.19]	[0.19]	(0.01)	
2-4 hours per week	0.10	0.10	0.00	2,237
	[0.30]	[0.30]	(0.02)	
4-6 hours per week	0.05	0.05	0.00	2,237
	[0.22]	[0.21]	(0.01)	
More than 6 hours per week	0.22	0.25	0.04*	2,237
	[0.41]	[0.43]	(0.02)	

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.14: Impact on item-response effort proxies (endline and follow-up)

	(1) Math	(2) Science	(3) Language
A. Endline			
Share of items attempted	0.02*	0.01	0.01**
	(0.01)	(0.01)	(0.00)
Control mean	0.96	0.96	0.98
Last item attempted	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
Control mean	0.99	0.99	1.00
N (students)	2496	2489	2423
B. Follow-up			
Share of items attempted	0.00	0.00	-0.01
	(0.00)	(0.00)	(0.00)
Control mean	0.97	0.98	0.97
Last item attempted	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
Control mean	0.99	1.00	1.00
N (students)	2361	2363	2359

Notes: This table shows the impact of the intervention on three item-response effort proxies constructed from assessments administered at endline, 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). The share of items attempted counts any response, including multiple selections, as an attempt. The last item attempted is the position of the final attempted item divided by the number of items in the assessment. The final measure is the share of the last five items attempted. Impact estimates come from regressions that include randomization-strata (i.e., grade) fixed effects and LASSO-selected baseline covariates. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.15: Impact on teacher workload, support, and self-efficacy (endline)

	(1) Control	(2) Treatment	(3) Difference	N (instructors)
A. Workload and support constraints				
Too many students in class	2.36	2.30	-0.06	91
	[0.90]	[1.07]	(0.21)	
Too much material to cover	2.83	2.66	-0.18	91
	[1.24]	[1.28]	(0.25)	
Too many teaching hours	2.45	2.23	-0.22	91
	[0.86]	[0.94]	(0.19)	
Needs more time to prepare	3.43	3.18	-0.25	91
	[1.18]	[1.30]	(0.24)	
Needs more time for individual help	3.79	3.64	-0.16	91
	[1.12]	[1.18]	(0.22)	
Too many administrative tasks	3.62	3.77	0.15	91
	[1.24]	[1.29]	(0.27)	
Lacks principal support	1.66	1.64	-0.03	91
	[0.94]	[0.84]	(0.13)	
Workload/support constraints index	0.00	-0.18	-0.19	91
	[1.00]	[1.10]	(0.20)	
B. Self-efficacy				
Can motivate low-interest students	4.26	4.11	-0.14	91
	[0.85]	[1.13]	(0.22)	
Can use alternative strategies	4.36	4.14	-0.22	91
	[0.90]	[1.07]	(0.23)	
Can motivate students to do well	4.36	4.11	-0.25	91
	[1.01]	[1.22]	(0.25)	
Can help families support students	3.96	3.80	-0.16	91
	[0.96]	[1.32]	(0.26)	
Can give alternative explanations	4.47	4.14	-0.33	91
	[0.88]	[1.23]	(0.25)	
Self-efficacy index	0.00	-0.27	-0.27	91
	[1.00]	[1.29]	(0.27)	

Notes: This table compares teachers' reports of workload, support, and self-efficacy in control and treatment classrooms, measured in endline teacher surveys 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). Higher values in panel A indicate more workload or support constraints; principal support is reverse-coded. Higher values in panel B indicate greater self-efficacy. The indexes are averages of item-specific z-scores, 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 A.16: Impact on allocation of time on task (announced observations)

	(1)	(2)	(3)	(4)	(5)
	Teachers		Impact on teachers	Fellows	Difference between teachers and fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time...					
...reading aloud	0.05	0.00	-0.05***	0.01	-0.04**
	[0.12]	[0.00]	(0.01)	[0.04]	(0.01)
...on explanation/lecture	0.34	0.02	-0.32***	0.34	-0.01
	[0.21]	[0.06]	(0.02)	[0.19]	(0.03)
...on interactive demo	0.04	0.00	-0.04***	0.01	-0.03***
	[0.07]	[0.00]	(0.01)	[0.03]	(0.01)
...on question and answers	0.18	0.00	-0.18***	0.15	-0.03
	[0.15]	[0.01]	(0.02)	[0.13]	(0.02)
...on practice and drill	0.01	0.00	-0.01*	0.00	0.00
	[0.03]	[0.01]	(0.00)	[0.02]	(0.01)
...on classwork	0.13	0.00	-0.13***	0.16	0.03
	[0.15]	[0.01]	(0.02)	[0.12]	(0.02)
...copying	0.03	0.00	-0.03***	0.06	0.03**
	[0.07]	[0.00]	(0.01)	[0.08]	(0.01)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Minutes of lesson time...					
...reading aloud	1.47	0.00	-1.47***	1.71	0.25
	[3.67]	[0.00]	(0.40)	[4.90]	(0.77)
...on explanation/lecture	10.3	2.40	-7.88***	40.4	30.1***
	[6.23]	[7.27]	(1.21)	[22.9]	(3.19)
...on interactive demo	1.09	0.00	-1.09***	0.98	-0.12
	[2.09]	[0.00]	(0.22)	[3.32]	(0.54)
...on question and answers	5.44	0.24	-5.20***	18.4	12.9***
	[4.59]	[1.70]	(0.51)	[15.1]	(2.01)
...on practice and drill	0.22	0.24	0.02	0.49	0.27
	[0.90]	[1.70]	(0.20)	[2.40]	(0.36)
...on classwork	3.88	0.24	-3.64***	18.9	15.0***
	[4.42]	[1.70]	(0.53)	[14.1]	(1.97)
...copying	0.97	0.00	-0.97***	6.86	5.89***
	[2.02]	[0.00]	(0.21)	[9.17]	(1.27)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of time on task of teachers in the control and treatment groups and of fellows in the treatment group, measured in announced observations about five months after the rollout of the intervention (December 2017). All lessons taught by teachers lasted 30 minutes and all lessons taught by 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.17: Impact on allocation of time on classroom management (announced observations)

	(1)	(2)	(3)	(4)	(5)
	Teachers		Impact on teachers	Fellows	Difference between teachers and fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time...					
...on instructions	0.01	0.01	-0.01	0.04	0.03***
	[0.04]	[0.02]	(0.01)	[0.08]	(0.01)
...on discipline	0.02	0.01	-0.01	0.02	0.00
	[0.04]	[0.04]	(0.01)	[0.05]	(0.01)
...on class mgmt. w/students	0.06	0.01	-0.05***	0.08	0.02*
	[0.08]	[0.03]	(0.01)	[0.08]	(0.01)
...on class mgmt. alone	0.06	0.18	0.13***	0.02	-0.04***
	[0.10]	[0.25]	(0.04)	[0.05]	(0.01)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Minutes of lesson time...					
...on instructions	0.34	0.72	0.38	5.14	4.80***
	[1.06]	[2.88]	(0.41)	[9.17]	(1.26)
...on discipline	0.44	1.20	0.76	1.96	1.52*
	[1.15]	[4.37]	(0.60)	[6.17]	(0.87)
...on class mgmt. w/students	1.72	0.96	-0.76	9.80	8.09***
	[2.53]	[3.29]	(0.58)	[9.06]	(1.31)
...on class mgmt. alone	1.66	22.1	20.4***	2.20	0.56
	[2.98]	[29.4]	(3.93)	[5.83]	(0.85)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of time on classroom management of teachers in the control and treatment groups and of fellows in the treatment group, measured in announced observations about five months after the rollout of the intervention (December 2017). All lessons taught by teachers lasted 30 minutes and all lessons taught by 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.18: Impact on allocation of time off task (announced observations)

	(1)	(2)	(3)	(4)	(5)
	Teachers		Impact on teachers	Fellows	Difference between teachers and fellows
	Control	Treatment	Col. (2)-(1)	Treatment	Col. (4)-(1)
A. Share of lesson					
Share of lesson time...					
...on social interaction w/students	0.01	0.00	-0.01	0.00	-0.01
	[0.03]	[0.01]	(0.00)	[0.01]	(0.00)
...on social interaction w/adults	0.02	0.02	0.00	0.01	-0.01
	[0.05]	[0.04]	(0.01)	[0.02]	(0.01)
...uninvolved	0.04	0.10	0.06*	0.02	-0.02
	[0.09]	[0.24]	(0.03)	[0.05]	(0.01)
...out of room	0.02	0.57	0.55***	0.07	0.05***
	[0.09]	[0.40]	(0.05)	[0.10]	(0.02)
N (instructors)	96	50	146	50	146
B. Time in lesson					
Minutes of lesson time...					
...on social interaction w/students	0.22	0.24	0.02	0.25	0.03
	[0.90]	[1.70]	(0.26)	[1.71]	(0.26)
...on social interaction w/adults	0.44	1.92	1.48**	0.74	0.30
	[1.38]	[4.44]	(0.66)	[2.91]	(0.42)
...uninvolved	1.13	12.2	11.1***	2.20	1.08
	[2.83]	[28.2]	(3.93)	[5.83]	(0.93)
...out of the room	0.72	68.6	67.9***	8.33	7.63***
	[2.57]	[47.4]	(6.27)	[12.3]	(1.72)
N (instructors)	96	50	146	50	146

Notes: This table compares the allocation of time off task of teachers in the control and treatment groups and of fellows in the treatment group, measured in announced observations about five months after the rollout of the intervention (December 2017). All lessons taught by teachers lasted 30 minutes and all lessons taught by 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.19: Impact on IRT-scaled test scores among high-scoring teachers (endline, endline audit, follow-up)

	(1) Math	(2) Science	(3) Language
A. Endline			
Treatment	0.33***	0.20***	0.16***
	(0.04)	(0.05)	(0.05)
N (students)	2067	2060	2007
B. Endline audit			
Treatment	0.13	0.24***	
	(0.09)	(0.07)	
N (students)	456	456	
C. Follow-up			
Treatment	0.35***	0.13**	0.05
	(0.04)	(0.05)	(0.04)
N (students)	1882	1883	1879

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 instructors whose overall IRT-scaled teacher-assessment scores fall within the common support of fellows and control-group teachers. Student IRT-scaled scores are standardized to have mean zero and standard deviation one in the control group. Impact estimates include randomization-strata fixed effects and LASSO-selected baseline covariates. Standard errors (clustered by school) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.20: Impact on class materials (announced observations)

	(1) Control	(2) Treatment	(3) Col. (2)-(1)	N (instructors)
Class has blackboard	0.98	1.00	0.02	150
	[0.14]	[0.00]	(0.02)	
Class has whiteboard	0.02	0.00	-0.02	150
	[0.14]	[0.00]	(0.02)	
Class has chalk/markers	0.98	1.00	0.02	150
	[0.14]	[0.00]	(0.02)	
Class has textbook for teachers	0.82	0.30	-0.52***	150
	[0.39]	[0.46]	(0.08)	
Class has textbook for students	0.72	0.26	-0.46***	150
	[0.45]	[0.44]	(0.10)	
Class has laptop	0.06	0.04	-0.02	150
	[0.24]	[0.20]	(0.05)	
Class has digital whiteboard	0.04	0.04	0.00	150
	[0.20]	[0.20]	(0.00)	
Class has LCD projector	0.08	0.06	-0.02	150
	[0.27]	[0.24]	(0.04)	
Class has TV	0.16	0.06	-0.10	150
	[0.37]	[0.24]	(0.07)	
Class has science/math equipment	0.30	0.16	-0.14*	150
	[0.46]	[0.37]	(0.08)	
Class has maps	0.24	0.12	-0.12	150
	[0.43]	[0.33]	(0.08)	
Class has charts/poster	0.76	0.78	0.02	150
	[0.43]	[0.42]	(0.08)	
Class has toys/games	0.08	0.04	-0.04	150
	[0.27]	[0.20]	(0.04)	
Composite index (std.)	0.00	-0.13	-0.13	150
	[1.00]	[0.71]	(0.10)	

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%.