Explaining the Productivity Paradox: Experimental Evidence from Educational Technology

Andreas de Barros[*]

September 27, 2022

Abstract

Explaining the productivity paradox—the phenomenon where an introduction of information and communication technology (ICT) does not lead to improvements in labor productivity—is difficult, as changes in technology often coincide with adjustments to working hours and substitution of labor. I conduct a cluster-randomized trial in India to investigate the effects of a program that provides teachers with continuous training and materials, encouraging them to blend their instruction with high-quality videos. Teaching hours, teacher-to-student assignments, and the curriculum are held constant. Eleven months after its launch, I document negative effects on student learning in grades 9 and 10 in mathematics, and no effects in science. I also find detrimental effects on instructional quality, instructional practices, and student perceptions and attitudes towards mathematics and science. These findings suggest adjustment costs can serve as one explanation for the paradox.

[*] I gratefully acknowledge the Government of Haryana and staff at Avanti Fellows for their contributions to the "Sankalp" project, and for their commitment to evidence-based education policy. I am very thankful for support by staff at The Abdul Latif Jameel Poverty Action Lab (J-PAL) South Asia office, especially Srijana Chandrashekhar, Anushka Ghosh, Brinda Sapra, and Anuja Venkatachalam. I would like to thank Abhijit Banerjee, Felipe Barrera-Osorio, Alejandro Ganimian, Andrew Ho, Isaac Mbiti, Karthik Muralidharan, Tavneet Suri, and Martin West for their feedback. I thank Heather Hill, Sharon Kim, Matthew Kraft, Ezequiel Molina, Edward Seidman, and Carmen Strigel for making instruments available. I thank the DISE team at the National University of Educational Planning and Administration (NUEPA) for making data available, and Emmerich Davies for sharing related resources. I thank J-PAL's Post-Primary Education (PPE) Initiative for funding. The Michael and Susan Dell Foundation supports Avanti Fellows with funding for the "Sankalp" project and its evaluation. This research is in compliance with Institutional Review Boards in India (at the Institute for Financial Management and Research) and the United States (at Harvard University and the Massachusetts Institute of Technology). The study has been pre-registered at the AEA RCT Trial Registry (AEARCTR-0003665). The usual disclaimer applies, and I have no conflicting interests to declare.

1 Introduction

The productivity paradox is concerned with the question of why rapid development of and investment in information and communication technology (ICT) does not always lead to increases in labor force productivity (Solow, 1987).[1] As education systems around the world invest in technological solutions to address low levels of student learning, a similar puzzle emerges: it is unclear whether such investments lead to improvements in educational productivity, and recent review articles point to both positive and negative impacts (Bulman and Fairlie, 2016; Escueta *et al.*, 2020).

Explaining and solving the productivity paradox is particularly difficult as technological investments often conflate changes in working hours, substitution of labor, *and* the potentially factor-augmenting role of technology. To isolate the causal effect of technology on individual labor productivity, one would need to hold working hours constant, maintain other labor inputs at a given level, and observe a plausibly exogeneous introduction of a technology-based production strategy. To focus on the effects of information and communication technology, one may also want to investigate whether the effects and cost-effectiveness of an ICT intervention systematically differ from those of another, low-tech (i.e., non-ICT) change in production technology.

Educational technology (EdTech) is a prime example of an ICT intervention whose potential complementarity with other supply-side factors remains largely unstudied. It also exemplifies the challenge of how to isolate the effect of technology on labor productivity. EdTech investments may lead to an increase in instructional inputs, including added instructional time.[2] Such investments may also substitute for instructional input factors.[3] Most commonly, however, the introduction of educational technology represents a mixture of changes in study time, substitution of teacher-led instruction, and adjustments to instructional technology.[4] Perhaps because of this challenge, educational technology has rarely been studied as a potential complement to teaching, and many studies have ignored the question of whether investments in EdTech indeed serve a purpose of factor-augmenting *technology*. In addition, it is unclear how the effectiveness of educational technology compares to that of other teacher-training solutions that also aim to improve teaching quality, yet do not require ICT.

[1] Solow's initial notion of a productivity paradox refers to the introduction of information and communication technology in the 1970s and 80s. More recently, with respect to artificial intelligence and machine learning, the phenomenon is also referred to as "modern" productivity paradox (see Brynjolfsson *et al.*, 2017).

[2] For example, students may access educational technology in an after-school program, or through distance education. For a recent evaluation of a technology-based intervention that is administered in an after-school program, see Muralidharan *et al.* (2019).

[3] For example, students may be taken to a computer-lab during class-time, where computer-based instruction replaces in-person instruction. For studies that seek to isolate substitution effects of computer-assisted learning, see Bettinger *et al.* (2020) and Ma *et al.* (2020). For studies of interventions that substitute in-person teaching with audio and video materials, see Fabregas (2018); Jamison *et al.* (1981); Johnston and Ksoll (2017); Naik *et al.* (2020); Navarro-Sola (2019); Seo (2017).

[4] For studies of interventions that substitute in-school teaching with one-on-one software that also replaces instructional content, see Araya *et al.* (2019); Banerjee *et al.* (2007); Carrillo *et al.* (2010); Lai *et al.* (2015); Linden (2008); Taylor (2018).

This article presents experimental evidence on the effects of a computer-assisted educational program that encourages teachers to blend their instruction with high-quality video materials. The program provides schools with infrastructure upgrades (including tablets for teachers and TVs), an application with video materials, accompanying workbooks, and related teacher training. A key characteristic of the program is that it *complements* a teacher's instruction—it does not seek to replace the teacher, nor does it add instructional time for students. Another notable feature is the program's alignment with the common curriculum of the schools it targets, and—uncommon for many technology solutions used in developing countries—it operates in vernacular language and does not require English. The intervention is also noteworthy for supporting teachers through continuous, on-site coaching in schools, beyond its initial off-site orientation. The program's delivery does not require internet availability, it does not need for students to have access to (or knowledge of) computers, and it is therefore less costly than other programs that call for such features.

I estimate the causal effects of the program through a randomized trial across 240 schools in eight districts of Haryana, India, and their grade-9 and grade-10 students ($n = 24,584$). To my best knowledge, this is the largest experiment on the effectiveness of computer-assisted instruction to date. Results are therefore estimated precisely. The study collaborates with a state government at substantial scale, and it operates in public schools, with government teachers, during the usual school hours. Hence, results may be influenced less by site-selection bias (Allcott, 2015) or by implementer effects (Vivalt, 2020). As discussed, other studies of educational technology are also often limited to investigating bundled interventions—in contrast, the present study teases out the effectiveness of computer-assisted instruction (in contrast to other program components), through separate experimental arms. My main outcome of interest is student learning in mathematics and science, as measured through paper-based tests, after approximately ten months of program implementation. Beyond test scores, I make use of detailed process-monitoring data, of student interviews, and of in-person classroom observations—as a result, the study goes beyond a mere "black-box" evaluation, providing granular information on program implementation, take-up, and potential mechanisms. My analyses of these data have been pre-registered—the study's findings are thus not prone to specification searching.

I begin my analyses by providing additional information on the study design and its validity. In a first step, I compare the study sample against rich, large-scale data for the universe of registered Indian public schools, and their locations. I find that study schools are positively selected into the sample within the state, but that their districts are representative for student performance in India. Next, I compare observable, time-invariant school and student characteristics for an experimental group of schools with the full information and communication technology program ("ICT schools"), an experimental group of schools that received the program without its technology-related components ("Workbook schools"), and an experimental group of Control schools that continued with "business-as-usual". I find that ICT and Control schools were statistically indistinguishable, before the program was rolled out, and introduce robustness checks to alleviate concerns that Workbook schools differed from the remaining two groups. Across the three groups I also find no differences in students' attrition rates. Finally, I show that the program was implemented as intended and taken up well, by providing information on teacher trainings, infrastructure upgrades, and program usage.

Thereafter, I present three sets of results. First, the study finds that, after about six months, students in schools assigned to the ICT intervention performed 0.16 standard deviations lower in mathematics, as compared to their peers in the comparison group with no intervention. Students in schools that received the program without the technology-related components ("Workbook schools") performed 0.07 standard deviations below the Control schools, but I cannot statistically distinguish their results from the remaining two groups. I do not find effects on student learning in science. The results suggest that these effects are largely uniform across cognitive domains (i.e., higher- vs lower-order thinking skills), across curricular grade-levels (i.e., at- vs below-level materials), and content domains (e.g., algebra vs geometry in mathematics, or biology vs chemistry in science).

Second, in analyses of heterogeneous effects, I find the negative effects in mathematics are largely uniform across students' grades. A non-parametric investigation of heterogeneous effects moreover shows that the impacts hold for a wide range of baseline performance levels. Finally, I find suggestive evidence for differences in the ICT program's effects across districts. There are few schools per district and estimates are more noisy. Keeping this caveat in mind, a comparison of the two most positively affected and the two most negatively affected districts suggest that impacts of the ICT intervention differed by 0.41 standard deviations in mathematics (with no difference in impacts for science).

Third, these results coincide with detrimental effects on two sets of potential mediators: observed instructional quality and student attitudes towards and perceptions of mathematics and science. Classroom observations document a reduction in the percentage of class time spent on instruction, for both treatment groups (of 6 and 7 percentage points, respectively). For ICT schools, I also find a large negative effect on a summary index of observed instructional quality (of 0.46 standard deviations). I do not find such an effect for schools in the Workbook group. One-on-one interviews moreover reveal that both treatment variants caused students to enjoy mathematics or science less, to find those subjects harder than other subjects, and to experience greater nervousness towards them. A summary index across these and other measures of student perceptions and attitudes documents a negative effect of 0.26 standard deviations for ICT schools, and of 0.24 standard deviations for Workbook schools.

The study and its findings contribute to a nascent body of literature on how to support instructional quality in places where teacher content knowledge is limited, by complementing classroom teaching with technological aides. Results from these studies are mixed. Beg *et al.* (2019) conduct a smaller pilot of a similar intervention in Pakistan, which is bundled with teacher training and an additional at-home tutoring component. They find positive effects among grade-8 students' performance in mathematics and science (of 0.2-0.3 standard deviations). Naslund-Hadley *et al.* (2014) investigate the impacts of an early-grade mathematics curriculum in Ecuador, which also includes an audio component along with the new curriculum, materials, and volunteers. They document positive effects of 0.16 standard deviations in test scores. Bai *et al.* (2016) study the effectiveness of computer-assisted instruction, in rural China. They find positive effects on grade-5 students' performance in English (of 0.08 standard deviations). Ferman *et al.* (2019) measure the effects of a program that promotes teachers' use of the "Khan Academy" software in their classes, in Brazil. Their results show improvements in students' attitudes towards mathematics, but no impacts on achievement, in grades 5 to 9. Berlinski and Busso (2017) use a small experiment in

Costa Rica to compare instruction with interactive whiteboards against other educational technology interventions, and a control group. They find negative effects of 0.17 standard deviations on grade-7 geometry scores.

By disentangling the effects of program components—blended instruction vs teacher training—this study also complements a smaller literature on teacher capacity building and in-service coaching. Academic reviews for developed countries (Fryer, 2017; Jackson *et al.*, 2014) and less-developed countries (Arancibia *et al.*, 2016; Evans and Popova, 2016; Bruns and Luque, 2014) point out that "traditional" teacher development is rarely evidence-based, and often inefficient or even detrimental, especially if implemented at scale (Kerwin and Thornton, 2021; Loyalka *et al.*, 2019; Zhang *et al.*, 2013). Instead, teacher development in the United States has therefore increasingly turned to a set of "alternative design features", such as job-embeddedness, on-site capacity building, repeat trainings (of greater intensity and duration), and feedback and coaching (Egert *et al.*, 2018; Kraft *et al.*, 2018; Lynch *et al.*, 2019). In less-developed countries, research on this type of teacher development is still largely inexistent, however, with only few exceptions (Castro *et al.*, 2019; Cilliers *et al.*, 2020; Bruns *et al.*, 2018; Majerowicz and Montero, 2018).

Finally, the results also add to a growing literature that investigates how interventions that provide additional inputs to schools and teachers can be made more effective. Educational technology interventions that simply add infrastructure and improve equipment (such as laptops or smart classrooms) have been found to be largely ineffective (see Escueta *et al.* (2020)). Beyond technology, similar observations have been made for interventions that provided textbooks (Glewwe *et al.*, 2009; Sabarwal *et al.*, 2014), flipcharts (Glewwe *et al.*, 2004), school improvement grants (Das *et al.*, 2013; Blimpo *et al.*, 2015), and increased teacher pay (de Ree *et al.*, 2018). A recent set of studies therefore seeks to answer the question of why additional teaching inputs often do not lead to learning gains, even in otherwise resource-constrained environments. Such research asks whether, to be effective, these inputs need to be bundled with complementary interventions (Barrera-Osorio *et al.*, 2018; Mbiti *et al.*, 2019).

The remainder of the article is organized as follows. Section 2 describes the study's context and provides intervention details. Section 3 discusses the evaluation design, including the study's data, sampling, randomization, analytical strategy, and sample characteristics, as well as implementation fidelity and program take-up. Section 4 provides results and Section 5 concludes.

2 The program

2.1 Context

The study takes place in Haryana, a state in Northern India with a population of 25.3m. In Haryana, more than 96 percent of youth in the 14-16 age group are still within the formal education system, both among boys and among girls (ASER, 2018).[5]

[5] Dhar *et al.* (2018) moreover confirm that these numbers do not only reflect enrollment, but also match actual attendance.

The study's student population faces high levels of poverty and marginalization. For example, in the study's school districts, more than 36 percent of secondary students do not have a literate mother, and less than 16 percent of students have a flush toilet at home (Dhar *et al.*, 2018). Moreover, 37 percent of Haryana's students belong to a "scheduled caste"—the lowest castes in India, which are officially regarded as socially disadvantaged (NAS, 2017). In 2017, Haryana's GDP per capita was approx. $2,800 (World Bank, 2018).

The study is being performed in collaboration with Haryana's State Government and its "Government Senior Secondary Schools" (GSSS).[6] These schools are predominantly rural (80 percent of schools), and teach an exclusively Hindi curriculum. In Section 3.2, I provide additional information on observable school characteristics and student performance—for the study sample, for Haryana, and India.

2.2 Intervention details

The intervention's main component encourages teachers to blend their instruction with high-quality video-based materials, as delivered through a "smart" TV set and a handheld tablet. The study's conceptualization of "blended instruction" thus follows Graham (2006, 5), who defines the term as a "combin[ation] of face-to-face instruction with computer-mediated instruction." The video materials support the given curriculum and they are used during the common school hours, by government teachers, during their regular classes. The intervention is therefore a complement; it does not substitute for teachers' usual instruction, nor does it add instructional time.

More specifically, the videos consist of short, self-contained recordings that are directly mapped to the official curriculum.[7] They are embedded in a tablet-based application, which organizes the materials along with the textbook's chapters and sub-chapters.[8] There are 1,127 videos in total; they are 2.5 minutes on average, they usually feature a presentation or an animation, and they are all in Hindi language (the common language of instruction).

To allow for the videos to be shown in class, schools also receive infrastructure upgrades. The program's goal is to provide each school with two working smart classrooms, two TVs, two tablets (with the software installed), and a power inverter. As the program relies on this infrastructure, it also requires a modification in time tabling, by changing the room allocation for the affected grades. Infrastructure upgrades began in February 2019 and the adjustments were completed by the beginning of the new school year, in July 2019.

The program's second component consists of the provision of printed workbooks for students. The workbooks are also aligned with the official curriculum. They provide additional explanations, remediation notes, and exercises, in Hindi language. Teachers are expected to use the workbook in class through a structured activity, during which students

[6] The study includes Government Senior Secondary Schools (GSSS) and Government Girls Senior Secondary Schools (GGSSS). For simplicity, I use "Government Senior Secondary Schools" (GSSS) to refer to both types of schools. As of 2016/17, 3,259 of Haryana's 7,782 senior secondary schools are GSSS (42 percent). The remaining schools are under private management.

[7] The schools follow a common Central Board for Secondary Education (CBSE) curriculum.

[8] Teachers may choose between two interfaces: One is designed to be more convenient for class planning, the other is intended to be used in-class.

exchange workbooks and engage in peer instruction. Students received their workbooks at the beginning of the school year (in July 2019).

Finally, the program provides in-service training to teachers, both off-site and on-site. After an orientation to principals (in February 2019), teachers received an initial off-site training, for two days, at the beginning of the school year (in July 2019). Thereafter, field staff visited schools throughout the school year.[9] During each visit, they record any infrastructure shortcomings in the school, observe classroom instruction (following a standardized rubric), and provide continuous feedback and on-site training to teachers. The staff-to-school ratio is approximately 1 to 16, and there are three additional supervisors.

The program was developed by a large Indian NGO ("Avanti Fellows") and it was implemented in partnership with Haryana's State Government. Appendix Table A1 summarizes the program components and provides additional details on the distribution of responsibilities across Avanti Fellows and the state government.

3 Evaluation design

3.1 Data

As detailed below, my primary data sources capture (a) implementation fidelity and program take-up, (b) teaching behaviors and instructional quality, (c) student perceptions and attitudes, and (d) student achievement. I further complement this information with (e) rich secondary data capturing village/town characteristics, school characteristics, student performance on state- and country-wide exams, and student demographics.

3.1.1 Implementation fidelity and program take-up

The study collects data on the program's three main components: Teacher training (off-site and on-site), ICT materials (infrastructure upgrades and Avanti videos) and their usage, as well as "low-tech" materials (Avanti workbooks) and their usage. I measure teachers' exposure to offsite trainings with sign-in sheets, during training events. I measure their exposure to onsite capacity-building activities through a tablet-based application whose completion is mandatory for Avanti staff, during school visits. Information on ICT infrastructure comes from an infrastructure audit conducted in December 2018 and from school visits.[10] I track ICT usage with fine-grained data from the software backend. I combine this information with ratings of teachers' usage of and familiarity with the ICT materials, from in-person classroom observations. Lastly, I measure the availability of Avanti's "low-tech" materials and their usage, through school visits, in-person classroom observations, and one-on-one student interviews.

[9] Staff members usually count with several years of work experience in the education sector, but they have not worked as teachers in Haryana's government schools as teachers.

[10] To avoid demand effects, in schools without the ICT intervention, questions on ICT infrastructure were paused at the beginning of the 2019 school year, and only reinstated in November 2019.

3.1.2 Teaching behaviors and quality of instruction

Teaching behaviors and the quality of instruction are assessed through two instruments: Classroom observations and student reports. During classroom observations, I administered a standard measure of time-on-task, instructional behaviors, use of instructional materials, and student involvement (a modified "Stallings Observation System"; see Stallings *et al.* (2014)). I also administered a novel classroom observation instrument to capture the quality of instructional practices students receive ("QUIP", for its acronym).[11] Trained observers thus rated the quality of instructional quality on a four-point scale, along six dimensions: Monitoring of student learning, quality of feedback, maximization of learning time, whether the classroom work is mathematically / scientifically dense, whether the presentation of content is clear and not distorted, and the level of richness of mathematics and science. I investigate the scores for each of these six dimensions but also generate a summary index, by calculating their inverse covariance-matrix-weighted average (following Anderson, 2008). In Appendix C, I provide additional information on the QUIP measure, including supporting validity evidence. During student interviews, I moreover administered a four-item battery of questions on instructional quality, which I adapted from the Trends in International Mathematics and Science Study's (TIMSS) item bank on teaching quality. I again report on answers to individual items and on their inverse covariance-matrix-weighted average.

3.1.3 Student perceptions and attitudes

I measure students' perceptions and attitudes towards mathematics and science through one-on-one interviews. More specifically, I adapted a five-item battery of questions, with a four-point scale, from the TIMSS Context Questionnaires' "measure of students positive affect toward mathematics and science". As with the previous measures of this study, I investigate answers to each of the individual questions but also generate a summary index, by calculating their inverse covariance-matrix-weighted average (following Anderson, 2008).

3.1.4 Student learning

The study's main outcome of interest is student learning in mathematics and science. I measure student learning with standardized assessments, which were administered as paper-based tests at baseline (in December 2018, when students were one grade below) and at follow-up (in November 2019, when students were enrolled in grades 9 and 10). Students were given two hours to complete each assessment.

I designed these tests to assess what students know and can do in four content domains of mathematics (algebra, geometry, number sense, and statistics) and three content domains of science (biology, chemistry, and physics). At follow-up, together with Avanti's subject-matter experts, I moreover classified test questions into two cognitive domains: measures of

[11] My instrument development greatly benefited from conversations with experts on classroom observation measures; in particular, Professors Heather Hill and Andrew Ho (of Harvard) and Sharon Kim and Edward Seidman (of NYU). I also thank Ezequiel Molina and the World Bank's SABER team for sharing their newly developed "TEACH" classroom observation instrument with me.

higher-order thinking skill ("HOTS") and measures of lower-order thinking skill ("LOTS"). About half of the test questions covered materials at students' grade level; the other half covered materials from up to two grade-levels below.[12]

I scaled the results using a two-parameter Item Response Theory (2PL IRT) model, separately for mathematics and science.[13] In doing so, I make use of repeated items to map students' performance at baseline and follow-up onto a common, continuous scale (Stocking and Lord, 1983). I also calculate separate IRT scores for students' performance on higher-order vs lower order thinking skills. I furthermore classify students into whether (or not) they have mastered mathematics and science materials at their enrolled grade-level, and below their grade-level. Similarly, I classify students into whether they are proficient in grade-level material for each of the seven content domains. These classifications rely on Cognitive Diagnostic Models (CDMs).[14] In Appendix D, I provide additional information on these student assessments, their properties, and on my psychometric approach.

3.1.5 Secondary data

I combine the above information with additional secondary data, in five steps. First, I include rich socio-economic information for each school's village or town (including from the most recent population census, economic census, and satellite-recorded night lights data). I do so by matching each school's geolocation to its village/town, using GIS information for India's 2011 census, and matching these villages/towns to data from the Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG) (Asher *et al.*, 2019). Second, I add detailed administrative records for all Indian government schools, as per the country's District Information System for Education (DISE), and match the study's schools to this data-set. Third, I compile district-level results for the country's most recent National Achievement Survey (NAS 2017). I do so by compiling information from district-wise report cards, for all of India, and matching the study's districts to this data-set. Fourth, I obtained school-level results for the state's 2017 grade-10 exit exams ("board exams"), from Haryana's Department of School Education. Finally, I match students to official enrollment rosters, to obtain their gender and date of birth.

3.2 Sampling of schools, representativeness

The study includes all ninth- and tenth-graders in 240 schools, in eight districts of Haryana. The sampling of schools followed a three-stage process. First, eight of Haryana's 22 districts were selected. Districts were chosen based on their number of Government Senior Secondary Schools, schools' level of proficiency, and districts' geographic proximity from each other. Next, 374 (out of 807 schools in these eight districts) Government Senior Secondary Schools were chosen for a school audit. For this audit, schools were chosen based on the availability and qualification of mathematics and science teachers. Schools also had to enrol at least

[12] Test items and their grade-level mapping relate to the official "CBSE/NCERT" school curriculum.

[13] See Jacob and Rothstein (2016) for an accessible introduction to Item Response Theory, in the economics literature.

[14] See de la Torre *et al.* (2016) for an accessible introduction to CDMs.

one student in a grade-11 science section. The audit identified 250 schools that counted with an appointed principal and with an additional room (which could be converted to a smart classroom); 240 of these schools were selected based on their principal's interest in the intervention.

Table 1 investigates whether the sample of schools is representative, by comparing it with all other public secondary schools in the state and in the country. Panel A shows how study schools are located in villages/towns of greater size (in hectares, and in terms of population size), both in comparison to other villages/towns in Haryana and when compared to the average Indian village/town. The study locations are also more highly developed (as measured by literacy rates, formal employment, non-agricultural employment, consumption, and night lights), and they count with more primary schools.

Panel B reports on school characteristics. Study schools are predominantly rural (80 percent), but slightly less so as compared to the remaining schools in the state (90 percent), and India (84 percent). They are also slightly larger, serve a greater percentage of male students, are less likely to be co-ed, and serve a greater percentage of students belonging to an "Other Backward Class" (OBC). They moreover employ a greater percentage of female teachers, and they employ more staff. Study schools are more likely to count with a computer-aided learning lab; however, their computer-per-student ratio is representative both for the state and for India.

Panel C focuses on the study's eight districts, and their students' performance on the National Achievement Survey (in 2017, for grade 8). Haryana performs below the remaining Indian districts (column (7)), but the study districts outperform the remaining state (column (9)). These two phenomena offset each other and the study districts are representative for India (column (8)).

Panel D directly compares students' performance on the state's board exams (in 2017, in grade 10). On average, students in study schools outperformed their peers elsewhere in the state. Unfortunately, results for the NAS and for board exams are not directly comparable. However, the positive selection within Haryana may roughly offset the difference between Haryana and the remaining country.

Taken together, study schools are positively selected according to village/town characteristics and according to observable school characteristics. Their students outperform those of other public schools in the state, but their districts' student performance is representative for India. Study schools may reflect student performance in India overall, but data limitations do not allow for a direct test.

3.3 Randomization

I randomly assigned the study's schools to three groups of 80 schools each–an Information and Communication Technology (ICT) Group, a Workbook Group, or a Control Group.

1. The **ICT Group** was assigned to receive the full program, which promotes blended instruction. This includes two "smart" classrooms with ICT infrastructure, digital content to supplement teaching instruction, printed practice workbooks for students,

Table 1: *Sample representativeness*

	Number of observations			Mean			Differences		
	India	Haryana	Sample	India	Haryana	Sample	Haryana vs Remaining India	Sample vs Remaining India	Sample vs Remaining Haryana
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Village/Town characteristics									
Total population	590874	6754	199	2047.72	3730.16	18975.53	1701.89***	16933.51***	15708.19***
				[40675.72]	[23457.88]	[75744.10]	(497.79)	(2883.83)	(1677.20)
Literate population (percentage)	590874	6754	199	0.57	0.63	0.67	0.05***	0.10***	0.04***
				[0.15]	[0.10]	[0.06]	(0.00)	(0.01)	(0.01)
Employed population (percentage)	538511	6578	197	0.06	0.05	0.08	-0.00***	0.03***	0.03***
				[0.09]	[0.07]	[0.11]	(0.00)	(0.01)	(0.01)
Share of households whose main source of income is cultivation	541623	6578	172	0.38	0.34	0.35	-0.04***	-0.03	0.01
				[0.29]	[0.21]	[0.18]	(0.00)	(0.02)	(0.02)
Rural mean per capita consumption	540227	6478	165	16474.40	21629.37	22773.47	5217.53***	6300.99***	1174.00***
				[5162.97]	[4259.03]	[3249.99]	(64.14)	(401.91)	(335.58)
Night light per grid cell (avg.)	572000	6826	199	5.54	13.89	16.09	8.45***	10.56***	2.27***
				[4.91]	[7.08]	[8.56]	(0.06)	(0.35)	(0.51)
Number of primary schools	583572	6604	167	1.42	1.69	2.88	0.27***	1.46**	1.22***
				[6.20]	[1.19]	[2.16]	(0.08)	(0.48)	(0.09)
Total Geographical Area (in Hectares)	583570	6604	167	419.31	628.00	1404.86	211.08***	985.83***	797.01***
				[2610.70]	[687.05]	[1652.22]	(32.31)	(202.05)	(52.95)
Panel B: School characteristics									
Rural school	74500	3259	240	0.84	0.90	0.80	0.06***	-0.05*	-0.11***
				[0.36]	[0.30]	[0.40]	(0.01)	(0.02)	(0.02)
School size, grades 7 and 8 (no. of students)	74500	3259	240	86.18	91.56	103.51	5.63**	17.39*	12.90**
				[118.50]	[70.18]	[81.17]	(2.12)	(7.66)	(4.70)
Female students (percentage)	67247	3221	240	0.50	0.49	0.42	-0.01**	-0.08***	-0.07***
				[0.24]	[0.29]	[0.34]	(0.00)	(0.02)	(0.02)
Percentage OBC	74500	3259	240	0.66	0.69	0.82	0.03***	0.16***	0.14***
				[0.39]	[0.29]	[0.25]	(0.01)	(0.03)	(0.02)
Total number of teachers	74500	3259	240	13.40	15.17	23.71	1.86***	10.35***	9.22***
				[10.20]	[8.67]	[8.49]	(0.18)	(0.66)	(0.56)
Female teachers (percentage)	74151	3258	239	0.38	0.40	0.46	0.02***	0.08***	0.06**
				[0.28]	[0.28]	[0.27]	(0.00)	(0.02)	(0.02)
School is co-ed (vs. single-sex)	74500	3259	240	0.87	0.80	0.76	-0.07***	-0.11***	-0.04
				[0.33]	[0.40]	[0.43]	(0.01)	(0.02)	(0.03)
Computer Aided Learning Lab	74500	3259	240	0.27	0.53	0.64	0.27***	0.37***	0.12***
				[0.44]	[0.50]	[0.48]	(0.01)	(0.03)	(0.03)
Computers / no. of students	63188	3221	240	0.15	0.24	0.23	0.09***	0.07	-0.01
				[0.93]	[0.26]	[0.20]	(0.02)	(0.06)	(0.02)
Panel C: District-level student performance (NAS)									
Average math score	670	21	8	41.07	36.31	38.63	-4.92*	-2.47	3.75*
				[9.02]	[3.88]	[3.62]	(1.99)	(3.21)	(1.57)
Average math score (female)	670	21	8	41.29	37.07	39.37	-4.36*	-1.95	3.72*
				[9.16]	[4.01]	[4.31]	(2.02)	(3.26)	(1.64)
Average math score (male)	670	21	8	40.79	35.45	37.69	-5.52**	-3.14	3.62*
				[9.16]	[4.24]	[3.92]	(2.02)	(3.26)	(1.77)
Average science score	670	21	8	42.95	40.93	43.16	-2.09	0.21	3.61**
				[8.99]	[3.53]	[3.06]	(1.99)	(3.20)	(1.40)
Average science score (female)	670	21	8	42.89	41.01	43.44	-1.94	0.56	3.93*
				[9.24]	[4.21]	[4.07]	(2.05)	(3.29)	(1.72)
Average science score (male)	670	21	8	42.99	40.84	42.90	-2.22	-0.09	3.32*
				[9.00]	[3.57]	[3.11]	(2.00)	(3.20)	(1.46)
Panel D: School-level student performance (board exams)									
Average score, overall		3254	240		38.13	42.00			4.18***
					[11.78]	[11.43]			(0.79)
Average score, math science		3254	240		38.57	41.88			3.57***
					[9.65]	[9.38]			(0.64)
Average score, math		3254	240		40.26	44.76			4.85***
					[11.89]	[12.02]			(0.79)
Average score, science		3254	240		36.89	39.00			2.28***
					[9.33]	[8.88]			(0.62)
Percentage failing, overall		3254	240		0.53	0.46			-0.08***
					[0.23]	[0.22]			(0.02)
Percentage failing, math science		3254	240		0.35	0.26			-0.09***
					[0.22]	[0.19]			(0.01)
Percentage above 50, overall		3254	240		0.43	0.50			0.08***
					[0.22]	[0.21]			(0.01)
Percentage above 50, math science		3254	240		0.41	0.49			0.10***
					[0.22]	[0.22]			(0.01)

Notes. This table provides descriptive statistics for India, Haryana, and the study sample. Village-/town characteristics match a school's geolocation to a polygon of village-/town boundaries from India's 2011 census. 2011 census data, India's 2013 economic census, and 2013 night lights data as per Asher *et al.* (2019). School characteristics as per U-DISE, including public secondary schools only. NAS refers to district-level eighth-grade performance in government schools as per the 2017 National Achievement Survey. Haryana board exam scores are from 2017 for tenth-grade government school students, aggregated to the school-level. Standard deviations in brackets; standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

and continued, on-site capacity building for mathematics and science teachers responsible for teaching the grade-9 and grade-10 curricula.

2. The **Workbook Group** (or “low-tech” group) was assigned to receive a partial variant of the program. Its components are equivalent to those administered in the previous group. However, *the group does not receive those particular components that promote blended instruction* (i.e., ICT-related infrastructure upgrades and digital content).

3. The **Control Group** was assigned to not receive the facilities, materials or training of the program. Its schools continued with “business-as-usual”.

To achieve similar control and treatment groups and to improve statistical power, randomization was stratified. Within districts, I sorted schools into randomization strata of three, based on their school-level results on Haryana’s state-level board exams. I randomized schools within these triplets. More specifically, I repeated this randomization procedure ten times, and selected the randomization with greatest statistical balance.[15]

Figure 1 provides an overview of the study’s geographic scope and the sample of schools, by treatment status.

3.4 Analytical strategy

I estimate the intent-to-treat effect (ITT) of the two different treatments on follow-up outcomes, using the following specification.

$$Y_{irs2} = \alpha_s + \sum_{k=1}^{2} \beta_{ks} T_{ki} + \boldsymbol{X}_{irs1} + \phi_r + \epsilon_{irs2} \tag{1}$$

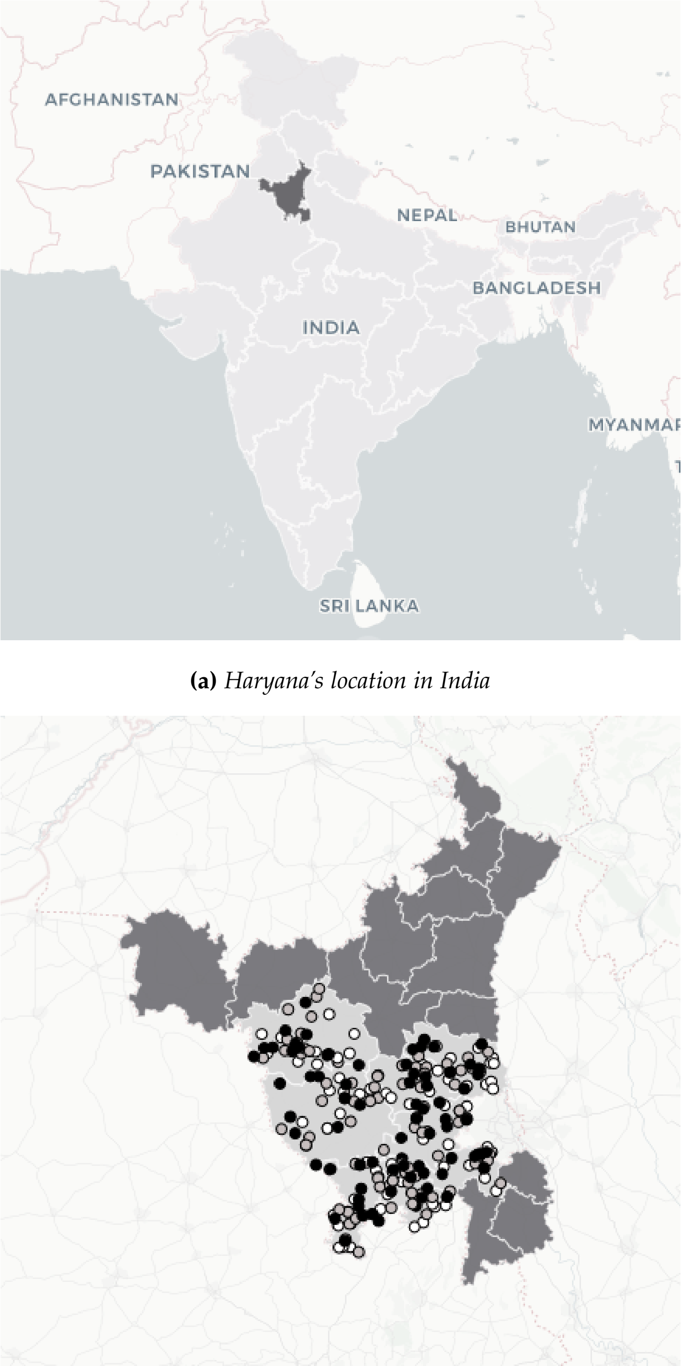
In Equation 1, Y_{irst} is the outcome of interest, for student i, in randomization stratum r, and subject s, at period t ($t = 1$ denotes baseline; $t = 2$ denotes follow-up). T_k is the dummy for treatment k. $\boldsymbol{X}_{irs1}$ is a vector of covariates measured at baseline; ϕ_r are randomization strata fixed effects and ϵ_{irs2} captures the idiosyncratic error term. Standard errors are clustered at the school-level (cf. Abadie *et al.*, 2017).

I select the vector of baseline controls through a LASSO procedure, following Dhar *et al.* (2018). For details, including a list of the selected controls, see Appendix E.

For the study’s main outcomes, in secondary analyses, I also use a specification that allows for heterogeneous treatment effects, by interacting potential moderators with the treatment

[15] To this end, I used LASSO to select a vector of covariates—from India’s District Information System for Education (DISE)—that were predictive of board exam results. Thereafter, I calculated t-statistics for board exam results and each of the selected variables (across the three experimental groups). I did so by estimating regressions of each characteristic on the treatment indicators and strata fixed effects. I then stored away the most extreme of these t-statistics, and selected the randomization where this value is smallest. See Bruhn and McKenzie (2009), who refer to this approach as “minmax method.” I am well aware that high numbers of re-randomization can lead to analytic problems, especially if the re-randomization strategy remains unknown. I follow Banerjee *et al.* (2017) by pre-specifing my strategy and choosing a conservative number of re-randomizations (ten re-randomizations).

Figure 1: *Geographic scope of the study*



(a) *Haryana's location in India*

(b) *Study districts and study schools, by treatment status*

Notes. Subfigure (a) shows the geographic location of Haryana in India. Subfigure (b) shows the study's eight selected districts in Haryana, and its 240 schools (by experimental group). ICT schools in black; Workbook schools in grey; Control schools in white.

indicators. I illustrate the corresponding specification for a sub-group analysis by grade, as follows.

$$Y_{irs2} = \alpha_s + \sum_{k=1}^{2} \beta_{ks} T_{ki} + \sum_{k=1}^{2} \beta_{2+ks} T_{ki} * G_{irs1} + \beta_5 G_{irs1} + \boldsymbol{X}_{irs1} + \phi_r + \epsilon_{irs2} \tag{2}$$

Here, G_{irs1} is the moderating variable of interest (in my illustration, an indicator for a student's grade), measured at baseline, and all else is defined as above. To avoid specification searching, I limit these analyses of heterogeneous effects to the following three moderators: Grade (as illustrated above), initial level of ability, and district.

In summary, my primary analyses thus assess the following research hypotheses, by testing their corresponding null, H.

1. The program's two variations affect student learning in subject s. H_1: in Equation 1, $\beta_{ks} \neq 0$
2. The two variants of the program affect student learning in subject s differently. H_2: in Equation 1, $\beta_{1s} \neq \beta_{2s}$

Respectively, my secondary analyses assess hypotheses of heterogeneous effects. They posit that the program's two variations affect student learning in subject s differently in grade 9 vs grade 10, that they have greater effects for weaker (/stronger) students, and that they differ by location (i.e., district). H_3: in Equation 2, $\beta_{2+ks} \neq 0$.

3.5 Balance and attrition

As shown in Table 2, randomization led to three groups of schools that are balanced in terms of observable, time-invariant school and student characteristics. Only one of 24 tests point to a difference in observable characteristics of schools' villages/towns. For Workbook school locations, inhabitants are slightly more likely to work in agriculture, as compared to Control school locations. Only 2 of 27 tests point to a difference in observable school characteristics. As compared to control schools, ICT schools are slightly more urban and Workbook schools serve a slightly greater percentage of students who belong to an "Other Backward Class". These differences do not go beyond what can be expected from multiple hypothesis testing. Moreover, none of the board exam results point to differences across the three groups. Students also do not differ in terms of their attrition rates, and student demographics are indistinguishable across groups (both at baseline, and among non-attritors).

Later in the paper, along with the study's program effects on student learning, I present balance checks on the baseline test in Table 4. As shown in Column (4), there are also no distinguishable differences across the ICT and Control groups on the baseline test. However, students in Workbook schools outperformed their peers in the Control group (by 0.19 standard deviations in mathematics, and 0.17 standard deviations in science) and in the ICT group (by 0.15 and 0.22 standard deviations, respectively). There is no consensus as to whether such baseline imbalance should be considered problematic (cf. Mutz *et al.*, 2018).

Table 2: *Observable, time-invariant school and student characteristics*

	Number of observations			Mean			Differences		
	Control	ICT	Workbook	Control	ICT	Workbook	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Village/Town characteristics									
Total population	79	79	80	51176.48	72475.19	74458.23	22455.55	-1656.66	24112.21
				[176477.19]	[200701.98]	[218836.33]	(22343.98)	(22238.33)	(22238.33)
Literate population (percentage)	79	79	80	0.67	0.68	0.68	0.00	-0.01	0.01
				[0.06]	[0.07]	[0.06]	(0.01)	(0.01)	(0.01)
Employed population (percentage)	78	78	80	0.10	0.11	0.11	0.00	-0.00	0.01
				[0.15]	[0.11]	[0.13]	(0.02)	(0.02)	(0.02)
Share of households whose main source of income is cultivation	66	58	65	0.36	0.36	0.32	-0.02	0.04	-0.06**
				[0.20]	[0.14]	[0.17]	(0.03)	(0.03)	(0.03)
Rural mean per capita consumption	63	55	61	22928.42	22854.26	22549.73	-51.31	381.96	-433.28
				[3375.93]	[3183.92]	[3137.32]	(630.48)	(630.48)	(616.62)
Night light per grid cell (avg.)	80	80	79	17.03	18.69	18.16	1.65	0.57	1.09
				[9.31]	[10.68]	[10.65]	(1.26)	(1.27)	(1.27)
Number of primary schools	63	56	62	2.75	3.20	2.82	0.48	0.39	0.09
				[2.09]	[2.71]	[1.87]	(0.46)	(0.46)	(0.45)
Total Geographical Area (in Hectares)	63	56	62	1195.79	1609.63	1572.68	356.07	252.86	103.20
				[886.55]	[1504.70]	[2266.09]	(219.64)	(216.54)	(213.39)
Panel B: School characteristics									
Rural school	80	80	80	0.84	0.74	0.81	-0.10*	-0.07	-0.02
				[0.37]	[0.44]	[0.39]	(0.06)	(0.06)	(0.06)
School size, grades 7 and 8 (no. of students)	80	80	80	111.47	101.42	97.64	-10.05	3.79	-13.84
				[96.21]	[69.95]	[75.36]	(10.00)	(10.00)	(10.00)
Female students (percentage)	80	80	80	0.44	0.39	0.44	-0.06	-0.05	-0.00
				[0.29]	[0.37]	[0.34]	(0.05)	(0.05)	(0.05)
Percentage OBC	80	80	80	0.78	0.82	0.85	0.04	-0.03	0.07*
				[0.28]	[0.25]	[0.22]	(0.03)	(0.03)	(0.03)
Total number of teachers	80	80	80	23.60	23.98	23.56	0.38	0.41	-0.04
				[8.78]	[7.86]	[8.89]	(1.21)	(1.21)	(1.21)
Female teachers (percentage)	80	80	79	0.45	0.44	0.48	-0.01	-0.03	0.03
				[0.24]	[0.29]	[0.27]	(0.03)	(0.03)	(0.03)
School is co-ed (vs. single-sex)	80	80	80	0.81	0.74	0.74	-0.07	0.00	-0.07
				[0.39]	[0.44]	[0.44]	(0.07)	(0.07)	(0.07)
Computer Aided Learning Lab	80	80	80	0.61	0.70	0.60	0.09	0.10	-0.01
				[0.49]	[0.46]	[0.49]	(0.07)	(0.07)	(0.07)
Computers / no. of students	80	80	80	0.22	0.24	0.23	0.02	0.01	0.01
				[0.18]	[0.20]	[0.23]	(0.03)	(0.03)	(0.03)
Panel C: School-level student performance (board exams)									
Average score, overall	80	80	80	42.04	42.01	41.95	-0.02	0.06	-0.08
				[11.55]	[11.62]	[11.27]	(0.30)	(0.30)	(0.30)
Average score, math science	80	80	80	41.89	41.75	42.00	-0.14	-0.24	0.11
				[9.45]	[9.31]	[9.49]	(0.53)	(0.53)	(0.53)
Average score, math	80	80	80	45.28	44.30	44.70	-0.98	-0.40	-0.58
				[12.09]	[11.84]	[12.25]	(1.13)	(1.13)	(1.13)
Average score, science	80	80	80	38.50	39.21	39.29	0.70	-0.09	0.79
				[9.07]	[9.06]	[8.58]	(0.64)	(0.64)	(0.64)
Percentage failing, overall	80	80	80	0.46	0.46	0.47	-0.00	-0.01	0.01
				[0.22]	[0.22]	[0.22]	(0.01)	(0.01)	(0.01)
Percentage failing, math science	80	80	80	0.26	0.27	0.27	0.01	0.01	0.01
				[0.18]	[0.19]	[0.20]	(0.01)	(0.01)	(0.01)
Percentage above 50, overall	80	80	80	0.50	0.50	0.50	-0.00	0.00	-0.00
				[0.21]	[0.21]	[0.21]	(0.01)	(0.01)	(0.01)
Percentage above 50, math science	80	80	80	0.50	0.48	0.50	-0.02	-0.02	0.00
				[0.22]	[0.23]	[0.22]	(0.02)	(0.02)	(0.02)
Panel D: Student characteristics									
Age (in years)	8601	8149	7699	14.27	14.25	14.28	-0.01	-0.04	0.03
				[1.21]	[1.23]	[1.24]	(0.04)	(0.04)	(0.04)
Female (%)	8601	8149	7699	0.48	0.50	0.48	0.02	0.03	-0.00
				[0.50]	[0.50]	[0.50]	(0.05)	(0.05)	(0.05)
Tested in follow-up	8665	8183	7736	0.75	0.76	0.76	-0.00	-0.01	0.01
				[0.43]	[0.43]	[0.43]	(0.01)	(0.01)	(0.01)
Panel E: Non-attritor characteristics									
Age (in years)	6536	6185	5895	14.15	14.14	14.16	-0.01	-0.04	0.03
				[1.14]	[1.17]	[1.17]	(0.04)	(0.04)	(0.04)
Female (%)	6536	6185	5895	0.50	0.52	0.51	0.02	0.01	0.00
				[0.50]	[0.50]	[0.50]	(0.05)	(0.05)	(0.04)

Notes. This table provides descriptive statistics for the study sample, by treatment status. Village-/town characteristics match a school's geolocation to a polygon of village-/town boundaries from India's 2011 census. 2011 census data, India's 2013 economic census, and 2013 night lights data as per Asher *et al.* (2019). School characteristics as per U-DISE. Village-/town characteristics and school characteristics are selected via LASSO (see Appendix E). Haryana board exam scores are from 2017 for tenth-grade government school students, aggregated to the school-level. Standard deviations in brackets; standard errors in parentheses (standard errors for individual-level data are clustered at the school level). All estimations include randomization strata fixed effects (F.E.s). * significant at 10%; ** significant at 5%; *** significant at 1%.

Yet, I address this issue with a pre-registered robustness check, in which the most severely imbalanced triplets of schools are dropped from the statistical analysis. In an iterative process, I drop randomization triplets until the mean differences across the three groups are below 0.05 standard deviations. This strategy suggests that seven triplets (and their 21 schools) need to be dropped, to achieve balance. The right-hand panels of Appendix Figure A1 show the results from this approach—by design, student performance becomes balanced across the three groups.

3.6 Implementation fidelity and take-up

For both treatment variants, the interventions were largely implemented as intended. As shown in Table 3, in either treatment group, teachers received both the initial off-site and subsequent on-site trainings (Panel A). Exposure to off-site training was only slightly higher in the ICT group, and the ICT group received more school visits, as compared to the treatment group (5.6 vs 3.5 visits, respectively).

Schools in the ICT group also successfully received the ICT infrastructure upgrades and Avanti's video materials. As shown in Appendix Table A2, all three groups started out with a large share of schools that counted with at least one smart classroom (more than 82 percent of schools; see Panel A). However, the ICT intervention added and repaired infrastructure in these rooms (Panel B), leading to large differences in the availability of functioning electricity, smart TVs, speakers, and tablets (Panel C). Less than 5 percent of Workbook and Control schools moreover counted with an ICT program that used a school's existing infrastructure (if functional), as compared to 100 percent of schools in the ICT group (Panel D).

As shown in Table 3's Panel B, these upgrades translated into substantial usage of the video materials, in both mathematics (534 minutes, on average) and science (755 minutes, on average). Importantly, the great majority (88 percent) of this usage occurred on days on which Avanti's field staff was not visiting a given school. During visits, Avanti staff reported video usage in about two thirds of classes (67 percent). They also marked less than a fifth of teachers for re-training (19 percent), and rated more than half of teachers (51 percent) as completely comfortable to navigate the software. Appendix Figure A2 provides additional detail on video usage over the study period.

Finally, schools in both treatment groups received the workbooks (95 and 99 percent, respectively), and showed usage of the same. During student interviews, about 9 out of 10 students reported having used the book, across both groups (89 percent), and more than 7 out of 10 students could produce the book when they were prompted (70 and 79 percent, respectively). For more than half of the students, their teacher had also started marking the workbook. During classroom observations, teachers in the Workbook group showed higher rates of usage (which may reflect their choice for video materials). Nevertheless, even in the Workbook group, only about a quarter of teachers used the materials "consistently" (25 percent) during class, conducted an Avanti in-class excise (27 percent), or allowed for student engagement during said exercise (20 percent).

Table 3: *Implementation fidelity and take-up*

	Follow-up mean		Difference (F.E.s)
	ICT	Workbook	ICT vs Workbook
	(1)	(2)	(3)
Panel A: Teacher training			
Teachers trained off-site (any)	3.75	3.21	0.54*
	[1.66]	[1.38]	(0.22)
Teachers trained off-site (mathematics)	1.59	1.32	0.26*
	[0.88]	[0.67]	(0.11)
Teachers trained off-site (science)	2.14	1.88	0.26
	[1.04]	[0.96]	(0.15)
Teachers trained off-site (grade 9 or 10)	3.55	3.10	0.45*
	[1.53]	[1.24]	(0.20)
Teachers trained off-site (grade 9)	3.05	2.85	0.20
	[1.37]	[1.23]	(0.18)
Teachers trained off-site (grade 10)	3.27	2.96	0.31
	[1.53]	[1.14]	(0.20)
On-site visits received	5.59	3.49	2.10***
	[2.03]	[0.98]	(0.22)
Panel B: Videos			
Usage (total, in min.)	1292.47		
	[845.94]		
Usage on days without visit (total, in min.)	1137.81		
	[798.27]		
Math usage (total, in min.)	537.93		
	[435.47]		
Math usage on days without visit (total, in min.)	465.53		
	[418.47]		
Science usage (total, in min.)	754.54		
	[541.12]		
Science usage on days without visit (total, in min.)	672.28		
	[508.85]		
Teacher showed Avanti video during obs.	0.67		
	[0.47]		
Teacher showed >1 type of video	0.55		
	[0.50]		
Teacher needs re-training	0.19		
	[0.40]		
Teacher comfortable to navigate independently	0.51		
	[0.50]		
Panel C: Workbooks			
Workbooks distributed	0.95	0.99	-0.04***
	[0.21]	[0.06]	(0.01)
Shortage of workbooks	0.16	0.22	-0.07*
	[0.36]	[0.25]	(0.03)
Student can produce the workbook when prompted	0.70	0.79	-0.11*
	[0.46]	[0.36]	(0.05)
Student has started using the workbook	0.89	0.89	0.09
	[0.31]	[0.27]	(0.06)
Student uses workbook in ’every class’	0.13	0.08	0.16**
	[0.33]	[0.26]	(0.06)
Workbook has been checked by a teacher	0.50	0.56	-0.00
	[0.50]	[0.49]	(0.18)
Workbook usage is ’consistent’	0.09	0.25	-0.17***
	[0.28]	[0.42]	(0.02)
Workbook usage is ’inconsistent’	0.30	0.37	-0.06*
	[0.46]	[0.47]	(0.03)
Workbook not used at all	0.62	0.37	0.23***
	[0.49]	[0.39]	(0.03)
Conducted in-class exercise	0.11	0.27	-0.15***
	[0.32]	[0.36]	(0.03)
Conducted in-class exercise, students involved	0.08	0.20	-0.11***
	[0.28]	[0.24]	(0.02)

Notes. This table provides descriptive statistics on program implementation and take-up, for the study sample’s treatment schools, by treatment status. “Follow-up” refers to all observations and student interviews conducted (between July 2019 and 31 December 2019). All estimations include randomization strata fixed effects (F.E.s). Standard deviations in brackets; standard errors in parentheses (standard errors for individual-level data are clustered at the school level). All estimations include randomization strata fixed effects (F.E.s). * significant at 10%; ** significant at 5%; *** significant at 1%.

Sample. Teacher training data for 160 treatment schools. Video usage data for 80 ICT schools. Data on familiarity with videos and workbook usage from 364 school visits, 1,015 classroom observations, and 916 student interviews, in treatment schools.

4 Results

4.1 Effects on student learning

Table 4 summarizes intent-to-treat (ITT) effects of the interventions on student learning. Panel A shows the study's main results. I find that, in mathematics, students of a school that was assigned to receive the full intervention (ICT schools) performed 0.16 standard deviations worse than their peers in a Control group school. Students in a school that was assigned to receive the intervention without its technology-related component (Workbook schools) performed 0.07 standard deviations worse than Control school students, but this coefficient is not statistically significant. Workbook students performed 0.09 standard deviations better than students in ICT schools, but the results cannot reject that the difference is zero. These effect sizes compare to 0.30 standard deviations of learning over the same time period, in the control group. Accordingly, the ICT intervention approximately halved students' year-to-year growth in mathematics. I do not find effects on science for either intervention.

Appendix Table A3 confirms these results are not influenced by two threats to the study's internal validity. First, recall that the study's randomization strategy was repeated ten times, to obtain a sample with three balanced groups. In Panel A of Appendix Table A3, I repeat the same re-randomization strategy, within each of 5,000 randomization iterations of a randomization inference-based analysis. The resulting p-value for the negative effect of the ICT intervention on mathematics is 0.04, and the overall results remain unchanged. Second, at baseline, average Workbook student outperformed students in the other two groups. Dropping the seven most imbalanced randomization strata from the analysis leads to more noisy results, but the study's substantive findings hold.

The remaining three panels document secondary results. They provide ITT effects on students learning by cognitive domains, curricular domains, and content domain. Panel B of Table 4 assesses whether the main impacts are driven by effects on higher-order vs lower-order thinking skills. Higher-order skills are captured by questions that require problem solving and knowledge transfer; in contrast, lower-order thinking skills are measured by questions that relate to procedural solutions and rote learning. Impacts on the continuous, standardized measures of these two cognitive domains are very similar.

Panel C provides information on whether students have "mastered" or are "proficient in" material at their enrolled grade-level, or below their enrolled grade-level. The negative effects in mathematics appear to be slightly larger for below-grade material, but this difference is not statistically significant. For the comparison of ICT schools again Workbook schools, the coefficient for at-grade-level science material becomes significant. Students in ICT schools were four percentage points less likely to have mastered these materials, in comparison to their peers in Workbook schools. Students in ICT schools were also four percentage points less likely to have mastered Science materials, as compared to students in Workbook schools.

Panel D shows the results for students' mastery on the different content domains measured by the test. In mathematics, because of the program, students in ICT schools were four percentage points less likely to have mastered algebra, three percentage points less likely to

Table 4: *ITT effects on student learning*

	Control group			Baseline differences (F.E.s)			Follow-up differences (F.E.s + Controls)		
	Baseline mean	Follow-up mean	Growth (F.E.s)	ICT vs Control	ICT vs Workbook	Workbook vs Control	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Effects on main outcomes									
Mathematics	0.07	0.37	0.30***	0.04	-0.15**	0.19**	-0.15**	-0.10	-0.06
	[1.00]	[1.04]	(0.07)	(0.07)	(0.08)	(0.08)	(0.06)	(0.06)	(0.06)
Science	0.07	0.15	0.08	-0.05	-0.21***	0.17**	-0.03	-0.05	0.02
	[0.99]	[0.93]	(0.06)	(0.07)	(0.07)	(0.08)	(0.05)	(0.05)	(0.05)
Panel B: Effects by cognitive domain									
Mathematics, higher-order		0.00					-0.14***	-0.09*	-0.05
		[1.00]					(0.05)	(0.05)	(0.05)
Mathematics, lower-order		0.00					-0.13**	-0.09	-0.04
		[1.00]					(0.06)	(0.06)	(0.06)
Science, higher-order		0.00					-0.03	-0.05	0.02
		[1.00]					(0.05)	(0.06)	(0.05)
Science, lower-order		0.00					-0.04	-0.07	0.03
		[1.00]					(0.05)	(0.06)	(0.05)
Panel C: Effects by curricular grade-level									
Mathematics, at grade-level	0.38	0.40	0.02	-0.01	-0.07**	0.06**	-0.04	-0.04	0.00
	[0.49]	[0.49]	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Mathematics, below grade-level	0.39	0.44	0.05*	0.00	-0.05*	0.05*	-0.06**	-0.03	-0.03
	[0.49]	[0.50]	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
Science, at grade-level	0.48	0.45	-0.03	-0.02	-0.07***	0.06**	-0.01	-0.04	0.03
	[0.50]	[0.50]	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Science, below grade-level	0.47	0.48	0.01	-0.01	-0.07***	0.06**	-0.02	-0.04**	0.03
	[0.50]	[0.50]	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Panel D: Effects by content domain									
Algebra	0.46	0.50	0.03*	0.01	-0.03*	0.04**	-0.04***	-0.03*	-0.02
	[0.26]	[0.27]	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
Geometry	0.50	0.45	-0.05***	0.01	-0.03**	0.04***	-0.03**	-0.02**	-0.01
	[0.22]	[0.25]	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number sense	0.43	0.45	0.02	-0.00	-0.04**	0.03*	-0.03**	-0.03*	-0.00
	[0.24]	[0.26]	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Statistics/Reasoning	0.49	0.46	-0.04**	0.01	-0.03**	0.03***	-0.02	-0.01	-0.01
	[0.21]	[0.24]	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Biology	0.45	0.50	0.05***	-0.01	-0.04***	0.03*	-0.01	-0.01	0.01
	[0.24]	[0.22]	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Chemistry	0.48	0.53	0.04***	-0.00	-0.05***	0.05***	-0.01	-0.01	0.00
	[0.23]	[0.23]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Physics	0.47	0.48	0.01	-0.02	-0.03**	0.02	-0.00	-0.01	0.01
	[0.22]	[0.22]	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)

Notes. This table presents the intent-to-treat (ITT) effects of the interventions on student learning. For reference, column (1) shows the baseline control group mean; columns (2) and (3) show the control group's growth from baseline to follow-up. "Baseline" refers to the assessment conducted in December 2018. "Follow-up" refers to the assessment conducted in November 2019. Panel A reports on test scores, as per a 2PL IRT model, and standardized to the baseline control group (including attritors). Panel B reports on test scores by cognitive domain (higher-order vs lower-order thinking skills), as per separate 2PL IRT models, and standardized to the follow-up control group (the necessary item mapping is not available for the baseline). Panels C and D report on the share of students having mastered the respective materials; mastery levels as per a Cognitive Diagnostic Model (see Appendix D). All estimations include randomization strata fixed effects (F.E.s). "Controls" indicates the inclusion of a vector of school- and village-level covariates, selected via LASSO (see Appendix E). Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.
Sample. 18,615 grade-9 and grade-10 students, in the study's 240 schools.

have mastered geometry and number sense, and two percentage points less likely to have mastered the statistics domain. In a comparison with their peers in Workbook schools, the negative effects of the ICT intervention are statistically significant for geometry and number sense. The remaining coefficients are not signficant at conventional levels.

4.2 Heterogeneity in treatment effects

In Table 5, I study whether the intent-to-treat effects on student learning differ by students' grade-level, their learning level at baseline, and by study district. Panel A suggests that the ICT program's negative effects in mathematics are similar across grade-9 and grade-10 students. Yet, a focus on grade-9 students also shows a statistically significant difference across students in the ICT and Workbook groups: In schools assigned to the ICT group, the grade-9 mathematics performance is 0.15 standard deviations below that of students in Workbook schools.

Panel B investigates differences in effects by students' performance on the baseline test. For both subjects and interventions, the point estimates appear slightly more negative for students who performed in the bottom two terciles, on the baseline test. However, this difference is not statistically significant. In Appendix Figure A3 I explore heterogeneity by student performance further, by non-parametrically plotting ITT effects against percentiles of baseline test scores. The above results are largely uniform across the range of student baseline performance. For science, the coefficients for both treatment variants are positive for approximately the top third of the distribution, but this "effect" is statistically indistinguishable from zero.

In Panel C, I explore the level of heterogeneity in treatment effects by study districts. I report the ITT effect in the two districts with the highest impact, the respective effect in the two districts with the most detrimental impact, and their difference. With 80 schools per treatment arm overall and eight districts in the study, these results should be interpreted with caution. With this caveat in mind, the results point to heterogeneity in the effects on mathematics learning, for the full intervention with the ICT component—the difference in ITT effects is 0.41 standard deviations. For science, and for the workbook-only intervention, the differences across districts are smaller and statistically indistinguishable from zero.[16]

4.3 Effects on potential mediators

In this section, I explore the effects of the program on two sets of potential mediators. I first report on impacts on instructional quality and instructional practices; thereafter, I report on impacts on student perceptions and attitudes.

[16] In Appendix Figure A4, I provide results for individual districts. As shown in the figure, a formal test supports that district-wise heterogeneity exceeds what could have been expected by chance, but it is difficult to identify individual districts that drive this heterogeneity. One district (Jhajjar) shows systematically better mathematics results, for a comparison of ICT schools with Workbook schools.

Table 5: *Heterogeneity in ITT effects on student learning*

	Control group			Baseline differences (F.E.s)			Follow-up differences (F.E.s + Controls)		
	Baseline mean	Follow-up mean	Growth (F.E.s)	ICT vs Control	ICT vs Workbook	Workbook vs Control	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: By grade									
Mathematics									
Grade 9	0.04	0.30	0.26***	-0.02	-0.14	0.12	-0.17**	-0.15**	-0.02
	[1.00]	[1.02]	(0.10)	(0.10)	(0.09)	(0.09)	(0.08)	(0.07)	(0.08)
Grade 10	0.09	0.41	0.32***	0.08	-0.15*	0.24***	-0.14*	-0.05	-0.08
	[1.01]	[1.04]	(0.08)	(0.09)	(0.09)	(0.09)	(0.07)	(0.07)	(0.07)
Grade 10 vs Grade 9	0.04	0.09*	0.06	0.10	-0.01	0.11	0.04	0.10	-0.06
	(0.08)	(0.05)	(0.10)	(0.12)	(0.10)	(0.10)	(0.08)	(0.08)	(0.08)
Science									
Grade 9	0.05	0.07	0.02	-0.11	-0.13*	0.02	-0.05	-0.09	0.04
	[1.00]	[0.90]	(0.03)	(0.09)	(0.08)	(0.10)	(0.06)	(0.06)	(0.07)
Grade 10	0.09	0.21	0.12***	0.01	-0.27***	0.28***	-0.02	-0.03	0.01
	[0.99]	[0.94]	(0.02)	(0.08)	(0.09)	(0.09)	(0.06)	(0.06)	(0.06)
Grade 10 vs Grade 9	0.03	0.14***	0.10***	0.12	-0.14	0.26**	0.03	0.07	-0.03
	(0.02)	(0.02)	(0.03)	(0.11)	(0.10)	(0.11)	(0.06)	(0.07)	(0.07)
Panel B: By baseline learning level									
Mathematics									
Bottom tercile	-0.96	-0.08	0.87***	0.04	0.03	0.01	-0.16**	-0.12*	-0.04
	[0.47]	[0.89]	(0.02)	(0.03)	(0.03)	(0.03)	(0.08)	(0.07)	(0.08)
Middle tercile	0.04	0.37	0.33***	-0.01	0.00	-0.02	-0.17**	-0.09	-0.09
	[0.26]	[0.94]	(0.02)	(0.02)	(0.02)	(0.02)	(0.07)	(0.06)	(0.06)
Top tercile	1.25	0.86	-0.39***	0.07	-0.08	0.15***	-0.11	-0.07	-0.04
	[0.59]	[1.06]	(0.03)	(0.05)	(0.05)	(0.06)	(0.09)	(0.08)	(0.09)
Top vs bottom tercile	2.05***	0.82***	-1.26***	0.03	-0.10	0.13*	0.05	0.05	-0.00
	(0.04)	(0.06)	(0.12)	(0.06)	(0.07)	(0.07)	(0.10)	(0.10)	(0.10)
Science									
Bottom tercile	-0.96	-0.25	0.72***	0.03	0.00	0.02	-0.07	-0.05	-0.02
	[0.57]	[0.88]	(0.02)	(0.03)	(0.04)	(0.04)	(0.07)	(0.07)	(0.07)
Middle tercile	0.07	0.14	0.07***	-0.01	-0.02	0.01	-0.05	-0.06	0.01
	[0.23]	[0.80]	(0.02)	(0.02)	(0.02)	(0.02)	(0.05)	(0.06)	(0.06)
Top tercile	1.18	0.59	-0.59***	-0.07	-0.17***	0.10	0.03	-0.05	0.09
	[0.56]	[0.91]	(0.02)	(0.05)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)
Top vs bottom tercile	1.99***	0.77***	-1.30***	-0.10	-0.18**	0.08	0.10	-0.01	0.11
	(0.04)	(0.05)	(0.09)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)
Panel C: By district									
Mathematics									
Two districts with highest effect	0.08	0.11	0.04	0.20	0.00	0.19	0.12	0.01	0.11
	[1.17]	[0.92]	(0.04)	(0.19)	(0.18)	(0.20)	(0.12)	(0.13)	(0.13)
Two districts with lowest effect	0.09	0.54	0.45***	-0.18	-0.12	-0.06	-0.43***	-0.26**	-0.17
	[1.09]	[1.12]	(0.04)	(0.12)	(0.13)	(0.11)	(0.13)	(0.11)	(0.16)
Highest-effect vs lowest-effect districts	-0.01	-0.43**	-0.42*	0.37*	0.12	0.26	0.55***	0.27	0.28
	(0.24)	(0.20)	(0.25)	(0.22)	(0.22)	(0.23)	(0.18)	(0.17)	(0.20)
Science									
Two districts with highest effect	-0.05	0.06	0.11***	-0.00	-0.32**	0.31*	0.10	0.08	0.01
	[0.99]	[0.96]	(0.04)	(0.18)	(0.15)	(0.17)	(0.15)	(0.14)	(0.13)
Two districts with lowest effect	0.18	0.34	0.16***	-0.13	-0.29**	0.16	-0.10	-0.09	-0.00
	[1.01]	[0.98]	(0.03)	(0.13)	(0.14)	(0.17)	(0.10)	(0.10)	(0.09)
Highest-effect vs lowest-effect districts	-0.23	-0.28	-0.05	0.12	-0.03	0.15	0.19	0.18	0.02
	(0.18)	(0.18)	(0.17)	(0.22)	(0.21)	(0.24)	(0.17)	(0.17)	(0.16)

Notes. This table presents on heterogeneity in intent-to-treat (ITT) effects of the interventions on student learning. For reference, column (1) shows the baseline control group mean; columns (2) and (3) show the control group's growth from baseline to follow-up. "Baseline" refers to the assessment conducted in December 2018. "Follow-up" refers to the assessment conducted in November 2019. All panels reports on test scores, as per a 2PL IRT model, and standardized to the baseline control group (including attritors). All estimations include randomization strata fixed effects (F.E.s). "Controls" indicates the inclusion of a vector of school- and village-level covariates, selected via LASSO (see Appendix E). Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.

Sample. 18,615 grade-9 and grade-10 students, in the study's 240 schools.

4.3.1 Instructional quality and teaching practices

The ICT program worsened the instructional quality students received. Panel A of Table 6 reports the ITT impacts on instructional quality, as measured by in-person classroom observations. As per the summary index, the ICT program led to a 0.46 standard deviation reduction in instructional quality (Columns 4 to 6). A comparison with students in Workbook school yields a negative effect of similar magnitude (0.30 standard deviation) and there is no overall effect of the Workbook intervention on instructional quality. In Appendix Table A4 and Appendix Table A5 I provide additional results by subject. I do not find substantial differences in these findings, across mathematics and science.

The ratings were given by the NGO that administers the program, which raises concerns about raters' impartiality. In Appendix C, I use external, video-based re-ratings of a subsample of classes, and find support for the hypothesis that the NGO-administered, in-person ratings are systematically higher in treatment classrooms.[17] In columns 7 to 9, I extrapolate this difference to all ratings. The adjusted findings suggest very large, negative impacts of the ICT program on instructional quality (of more than one standard deviations). Appendix Table A4 and Appendix Table A5 suggest that these effects are driven by impacts in mathematics. For the workbook-only intervention, I find negative effects in mathematics (of 0.52 standard deviations) and positive effects in science (of 0.36 standard deviations). In the following, I will focus on the ratings as provided by the NGO, but note that—for the ICT intervention—the findings should be considered an upper bound of true effects.

A breakdown of impacts by sub-dimensions of instruction detects the ICT program's negative effect across all areas of instructional quality. Observers rated the instruction to be of lower quality in terms of teachers' monitoring of student learning (0.32 standard deviations) and the quality of feedback students received (0.25 standard deviations). The program also negatively affected learning time (0.46 standard deviations) and the extent to which classroom work was perceived to be densely focused on mathematics / science (0.43 standard deviations). Moreover, I find detrimental effects on the presentation and quality of content (0.32 standard deviations) and the level of richness or depth of instruction (0.25 standard deviations). Conversely, I find that the Workbook intervention shifted the quality of instruction differentially across dimensions. While instructional density, the quality of content, and richness decreased (by 0.35, 0.32, and 0.21 standard deviations, respectively), teachers' level of monitoring and the quality of feedback to students may have improved (by 0.11 and 0.10 standard deviations, not significant).

In Panel B, I show the program's effects on observed instructional practices. These findings report on effects at the extensive margin of instruction, and they also provide additional information on immediate outputs (complementing the article's previous discussion of implementation fidelity). Both variants of the program led to a reduction of instructional time. In ICT schools, teachers spent nine percentage points less time teaching, and seven percentage points more time on off-task activities. In Workbook schools, a seven percentage-point reduction in instructional time coincided with a nine percentage-point

[17] This finding may reflect bias. However, it could also reflect that video-based ratings do not capture the same aspects of instruction.

Table 6: *ITT effects on instructional quality and teaching practices*

	Follow-up (most recent school visit)			Follow-up differences (F.E.s + Controls)			Follow-up differences (F.E.s + Controls + Adjust.)		
	Control	ICT	Workbook	ICT vs Control	ICT vs Workbook	Workbook vs Control	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Effects on observed instructional quality									
Monitoring of student learning	0.03	-0.23	0.03	-0.32***	-0.44***	0.11	-0.78***	-0.52***	-0.26***
	[0.99]	[1.03]	[1.06]	(0.10)	(0.09)	(0.10)	(0.10)	(0.10)	(0.09)
Feedback	0.03	-0.20	0.13	-0.25***	-0.36***	0.10	-0.44***	-0.61***	0.17*
	[0.99]	[0.92]	[1.14]	(0.09)	(0.11)	(0.11)	(0.08)	(0.09)	(0.09)
Management of class time	0.15	-0.21	-0.18	-0.46***	-0.11	-0.35***	-1.04***	-0.51***	-0.53***
	[0.93]	[1.24]	[1.12]	(0.11)	(0.12)	(0.11)	(0.11)	(0.12)	(0.10)
Dense focus on math/science	0.08	-0.34	-0.27	-0.43***	-0.10	-0.32***	-1.00***	-0.91***	-0.09
	[0.99]	[1.23]	[1.18]	(0.12)	(0.12)	(0.11)	(0.12)	(0.12)	(0.11)
Clarity, lack of errors	0.08	-0.10	-0.10	-0.32***	-0.15	-0.17	-0.48***	-0.65***	0.17
	[1.00]	[1.18]	[1.15]	(0.11)	(0.12)	(0.11)	(0.11)	(0.11)	(0.10)
Richness	-0.01	-0.23	-0.18	-0.25**	-0.04	-0.21**	-1.51***	-1.09***	-0.42***
	[0.99]	[0.88]	[0.93]	(0.10)	(0.10)	(0.10)	(0.09)	(0.09)	(0.08)
QUIP (Index)	0.09	-0.26	-0.07	-0.46***	-0.30***	-0.16	-1.14***	-1.04***	-0.10
	[1.00]	[1.05]	[1.15]	(0.11)	(0.11)	(0.12)	(0.11)	(0.11)	(0.10)
Panel B: Effects on observed teaching practices									
Instruction (% of class time)	0.64	0.56	0.60	-0.11***	-0.04	-0.07**	-0.09***	-0.02	-0.07**
	[0.32]	[0.32]	[0.33]	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Management (% of class time)	0.06	0.10	0.07	0.04***	0.03***	0.01	0.04***	0.04***	0.01
	[0.12]	[0.14]	[0.11]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Off-task (% of class time)	0.30	0.33	0.33	0.07**	0.01	0.06*	0.05	-0.01	0.06*
	[0.33]	[0.35]	[0.34]	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Class held in smart classroom (% of classes)	0.00	0.65	0.00	0.67***	0.65***	0.02	0.66***	0.63***	0.03
	[.]	[0.48]	[.]	(0.04)	(0.04)	(0.02)	(0.04)	(0.04)	(0.02)
Use of ICT (% of classes)	0.00	0.57	0.00	0.56***	0.55***	0.01	0.56***	0.54***	0.01
	[.]	[0.50]	[.]	(0.04)	(0.03)	(0.02)	(0.04)	(0.03)	(0.02)
Use of ICT (% of class time)	0.00	0.34	0.00	0.32***	0.31***	0.01	0.32***	0.30***	0.01
	[.]	[0.34]	[.]	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)
Use of textbooks (% of classes)	0.40	0.14	0.49	-0.24***	-0.30***	0.06	-0.24***	-0.30***	0.07
	[0.49]	[0.34]	[0.50]	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Use of textbooks (% of class time)	0.16	0.05	0.20	-0.10***	-0.14***	0.04	-0.10***	-0.14***	0.03
	[0.25]	[0.15]	[0.27]	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Use of notebooks (% of classes)	0.19	0.13	0.29	-0.07*	-0.13***	0.06	-0.07*	-0.15***	0.07*
	[0.39]	[0.33]	[0.45]	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)
Use of notebooks (% of class time)	0.07	0.05	0.10	-0.03	-0.05***	0.02	-0.04**	-0.06***	0.02
	[0.18]	[0.14]	[0.20]	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Group activity (% of classes)	0.01	0.00	0.01	-0.01	-0.01	0.00	-0.00	-0.01	0.00
	[0.09]	[0.07]	[0.11]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Group activity (% of class time)	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00
	[0.02]	[0.01]	[0.03]	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Panel C: Effects on student-reported teaching practices									
Student easily understands what the teacher teaches	-0.03	0.08	0.04	0.04	0.08	-0.04	0.07	0.07	0.01
	[1.01]	[0.96]	[0.96]	(0.09)	(0.08)	(0.09)	(0.09)	(0.07)	(0.08)
Teacher gives interesting things to do in class	0.05	0.08	-0.06	-0.12	0.08	-0.20**	-0.06	0.05	-0.11
	[0.99]	[0.96]	[1.06]	(0.10)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)
Teacher explains topic again if students do not understand	0.04	-0.06	0.10	-0.07	-0.13	0.06	-0.14	-0.13	-0.01
	[0.92]	[1.18]	[0.85]	(0.09)	(0.08)	(0.10)	(0.08)	(0.08)	(0.10)
Teacher does a variety of things to help learn	0.05	0.08	0.05	0.04	0.07	-0.03	0.05	0.10	-0.06
	[0.97]	[0.99]	[0.94]	(0.09)	(0.08)	(0.09)	(0.08)	(0.07)	(0.08)
Index	0.04	0.07	0.05	-0.04	0.03	-0.08	-0.03	0.03	-0.06
	[1.00]	[0.99]	[1.00]	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Teacher used videos to teach, past week	0.03	0.58	0.01	0.59***	0.59***	-0.00	0.60***	0.59***	0.00
	[0.16]	[0.49]	[0.10]	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Student usally works with at least one peer	0.87	0.83	0.79	-0.01	0.05	-0.06	-0.01	0.06**	-0.07**
	[0.33]	[0.38]	[0.41]	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)
Student usually works in groups	0.46	0.39	0.38	-0.15**	-0.01	-0.14**	-0.07	-0.01	-0.06
	[0.50]	[0.49]	[0.49]	(0.06)	(0.06)	(0.06)	(0.05)	(0.04)	(0.06)
# of math / science classes, past week	5.36	5.05	4.93	-0.28	0.17	-0.44	-0.22	0.27*	-0.49**
	[1.67]	[1.87]	[1.83]	(0.26)	(0.24)	(0.27)	(0.19)	(0.15)	(0.20)

Notes. This table presents the intent-to-treat (ITT) effects of the interventions on lesson-level measures of instructional quality and teaching practices. "Index" refers to the inverse covariance matrix-weighted average, following Anderson (2008). "Follow-up" refers to data from the most recent school visit, conducted in October-November 2019. All estimations include randomization strata fixed effects (F.E.s). All estimations include all observations (not just those from the most recent school visit) and interact treatment effects with a linear time trend; the reference date is the most recent observation date (November 15, 2019). "Controls" indicates the inclusion of a vector of school- and village-level covariates, selected via LASSO (see Appendix E). In Panel A, "adjust." refers to an adjustment for systematic differences in comparison to video-based re-ratings (see Appendix C), and the inclusion of rater fixed effects; in Panels B and C, "adjust." refers to the inclusion of rater fixed effects. Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.

Sample. Panel A and Panel B from 1,343 classroom observations in mathematics and science. Panel C from 1,214 student interviews. School visits follow a random schedule, classrooms and students are subsampled at random, and interview questions about subjects (mathematics vs science) are assigned at random.

increase in off-task activities. In ICT schools, teachers also spent three percentage points more time on classroom management, as compared to the Workbook group.

The classroom observations also confirm how teachers in ICT schools moved their instruction to smart classrooms (a 70 percentage-point increase) and increased their usage of (any type of) ICT-supported materials during classes (a 54 percentage point increase in any usage during class, and a 30 percentage point increase in the time spent using such materials). At the same time, the ICT program appears to have replaced teachers' usage of textbooks and notebooks. Moreover, the workbook-only version of the program did not lead to an increase of instruction with textbooks or notebooks. This may suggest that the program's workbooks were not used during class time. However, it may also suggest that workbooks either replaced and complemented existing usage of other textbooks and notebooks. Lastly, group work is hardly ever observed in classes, and the program did not change this practice.

Panel C complements these findings with information from student interviews. I do not find effects on the index of student-reported quality of instruction.[18] The remaining results confirm an increase in teachers' use of videos, in the ICT group. They also document how students in the Workbook group engaged in collaborative classroom work less often (a difference of seven percentage points). Finally, the results of Panel C suggest a slight decrease in the number of mathematics and science classes, for both treatment groups (for the ICT group, the difference is not statistically significant).

4.3.2 Student perceptions and attitudes

Both variants of the program led to negative effects on student perceptions and attitudes towards mathematics and science. Table 7 summarizes results from the study's one-on-one student interviews. Students in both treatment groups reported to enjoy mathematics and science less, to experience greater nervousness towards these subjects, and to find them harder than other subjects. Coefficients for the remaining two questions are negative as well, but statistically insignificant. The overall index of student perceptions and attitudes documents a negative impact of 0.26 standard deviations for the ICT intervention, and of 0.24 standard deviations for the intervention without the ICT component.

Appendix Table A6 repeats the above analysis by subject.[19] Its findings suggest that the negative program effects are concentrated in mathematics. In comparison to the Control group, for mathematics, the overall index of student perceptions and attitudes is 0.44 standard deviations lower among students in the ICT group, and 0.37 standard deviations lower among students in the Workbook group. In contrast, for science, the respective effects are substantially smaller and not statistically distinguishable from zero.

[18] Considering the limited predictive validity of student-reported instructional quality (Bacher-Hicks *et al.*, 2019), I place less emphasis on this finding. Accordingly, the study's pre-analysis plan defined classroom observations as main measure of instructional quality.

[19] At random, student interviews asked about perceptions and attitudes towards mathematics *or* science.

Table 7: *ITT effects on student perceptions and attitudes towards mathematics and science*

	Follow-up (most recent school visit)			Follow-up differences (F.E.s + Controls)			Follow-up differences (F.E.s + Controls + Adjust.)		
	Control	ICT	Workbook	ICT vs Control	ICT vs Workbook	Workbook vs Control	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
I enjoy learning [math/science]	0.03	-0.09	-0.05	-0.16*	-0.01	-0.15*	-0.20**	-0.01	-0.19**
	[1.05]	[0.94]	[0.94]	(0.09)	(0.07)	(0.09)	(0.08)	(0.07)	(0.09)
I learn many interesting things in [math/science]	0.01	-0.05	-0.02	-0.07	-0.02	-0.05	-0.11	-0.02	-0.09
	[1.01]	[0.94]	[0.98]	(0.09)	(0.08)	(0.09)	(0.07)	(0.08)	(0.08)
[Math/Science] makes me nervous (reversed)	-0.01	-0.05	-0.08	-0.18*	-0.02	-0.15	-0.23***	-0.01	-0.22**
	[1.01]	[0.95]	[0.94]	(0.09)	(0.08)	(0.09)	(0.08)	(0.08)	(0.09)
[Math/Science] is harder than other subjects (reversed)	0.03	-0.02	-0.05	-0.14*	0.02	-0.16*	-0.17**	0.04	-0.21***
	[1.00]	[0.99]	[0.99]	(0.09)	(0.08)	(0.09)	(0.08)	(0.07)	(0.08)
I don't understand what is taught in [math/science] (reversed)	0.06	0.05	0.07	-0.10	-0.04	-0.06	-0.13	-0.03	-0.10
	[1.01]	[1.04]	[1.10]	(0.10)	(0.08)	(0.10)	(0.09)	(0.08)	(0.09)
Index	0.04	-0.05	-0.04	-0.20**	-0.02	-0.17*	-0.26***	-0.01	-0.24***
	[1.02]	[1.01]	[1.04]	(0.10)	(0.08)	(0.10)	(0.08)	(0.08)	(0.09)

Notes. This table presents the intent-to-treat (ITT) effects of the interventions on student-level perceptions and attitudes towards mathematics and science. "Index" refers to the inverse covariance matrix-weighted average of the five questions, following Anderson (2008). "Follow-up" refers to data from the most recent school visit, conducted in October-November 2019. All estimations include randomization strata fixed effects (F.E.s). All estimations include all observations (not just those from the most recent school visit) and interact treatment effects with a linear time trend; the reference date is the most recent observation date (November 15, 2019). "Controls" indicates the inclusion of a vector of student, school-, and village-level covariates, selected via LASSO (see Appendix E). "Adjust." refers to the inclusion of interviewer fixed effects. Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.
Sample. 1,214 student interviews. School visits follow a random schedule, students are subsampled at random, and interview questions about subjects (mathematics vs science) are assigned at random.

5 Conclusion

I present experimental evidence on the short-term impacts of a computer-assisted educational program that encourages teachers to blend their instruction with high-quality video materials. I find that the program—which provides teachers with infrastructure upgrades, workbooks, continuous capacity building, and video materials—led to negative effects on mathematics test scores, and that it had no effects on student achievement in science. For the two subjects, these effects are similar across cognitive domains, across curricular grade-levels, and content domains. I find suggestive evidence for slightly larger (that is, more detrimental) effects among grade-9 (vs grade-10) students, but the findings are otherwise largely uniform across a wide range of students' baseline performance levels.

In my opinion, these findings reflect the program's negative impacts on instructional quality and practices. The program also led to worsened student perceptions and attitudes, in particular towards mathematics. Of course, it is impossible to establish the full mediating pathway causally, and additional intermediary outcomes may be at play as well. However, it is notable that the program's detrimental effects on these factors *coincided* with its effects on test scores. At the same time, the findings do not reflect implementation failure. The study's fine-grained data show how the intervention was implemented well, and how it led to substantial program usage in schools.

Taken together, the results shed light on the mechanisms behind the productivity paradox. They document how—at least in the short run—the introduction of information and communication technology does not always lead to improved labor productivity, including in a context where working hours, labor, and tasks are held constant. Rather, they point to disruptive effects, as workers struggle to adjust to a new technology. For the education sector, the results may be best interpreted as a cautionary warning that interventions that aim to improve instructional quality through technological aids come with adjustment costs, and may not lead to immediate improvements in student learning, even if they are implemented well.

The interpretation of results nevertheless requires some level of additional caution. They reflect impacts after only approximately one year of program implementation, and most students were only exposed to the program for about five months. The program's effectiveness may increase over time. The results may also not be entirely attributed to the program's video-based materials. Results for comparisons of schools that received the full intervention with a separate treatment group (that did not receive the program components related to educational technology) paint a complex picture. Overall, effects on test scores are indistinguishable across the two group, but the negative effects in grade 9 are larger in the ICT group. Instructional quality reduced in ICT schools only, not in schools without the technology component, but students perceptions and attitudes worsened in both groups of schools. Lastly, additional research is needed to better understand the difference in effects across mathematics and science.

References

Abadie, A., Athey, S., Imbens, G. and Wooldridge, J. (2017). *When Should You Adjust Standard Errors for Clustering?* Tech. Rep. w24003, National Bureau of Economic Research, Cambridge, MA.

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics*, **130** (3), 1117–1165.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y. and Lun, J. (2011). An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement. *Science*, **333** (6045), 1034–1037.

American Educational Research Association (ed.) (2014). *Standards for Educational and Psychological Testing*. Washington, D.C: American Educational Research Association.

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, **103** (484), 1481–1495.

Arancibia, V., Popova, A. and Evans, D. K. (2016). *Training Teachers on the Job: What Works and How to Measure it*. Working Paper 7834, The World Bank, Washington, D.C.

Araya, R., Arias Ortiz, E., Bottan, N. L. and Cristia, J. P. (2019). *Does Gamification in Education Work? Experimental Evidence from Chile*. Working Paper IDB-WP-982, Inter-American Development Bank, Washington, D.C.

Asher, S., Lunt, T., Matsuura, R. and Novosad, P. (2019). The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG), working paper.

Bacher-Hicks, A., Chin, M. J., Kane, T. J. and Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, **73**, 101919.

Bai, Y., Mo, D., Zhang, L., Boswell, M. and Rozelle, S. (2016). The impact of integrating ICT with teaching: Evidence from a randomized controlled trial in rural schools in China. *Computers & Education*, **96**, 1–14.

Banerjee, A., Chassang, S. and Snowberg, E. (2017). Decision Theoretic Approaches to Experiment Design and External Validity. In A. V. Banerjee and E. Duflo (eds.), *Handbook of Economic Field Experiments*, vol. 1, Elsevier, pp. 73–140.

Banerjee, A. V., Cole, S., Duflo, E. and Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*, **122** (3), 1235–1264.

Barrera-Osorio, F., García, S., Rodríguez, C., Sánchez, F. and Arbeláez, M. (2018). Concentrating Efforts on Low-Performing Schools: Impact Estimates from a Quasi-Experimental Design. *Economics of Education Review*, **66**, 73–91.

Beg, S. A., Lucas, A. M., Halim, W. and Saif, U. (2019). *Beyond the Basics: Improving Post-Primary Content Delivery through Classroom Technology*. Working Paper 25704, National Bureau of Economic Research.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, **28** (2), 29–50.

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

Bettinger, E., Fairlie, R., Kapuza, A., Kardanova, E., Loyalka, P. and Zakharov, A. (2020). *Does EdTech Substitute for Traditional Learning? Experimental Estimates of the Educational Production Function*. Tech. Rep. w26967, National Bureau of Economic Research, Cambridge, MA.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley, pp. 397–479.

Blimpo, M. P., Evans, D. and Lahire, N. (2015). *Parental Human Capital and Effective School Management: Evidence from the Gambia*. Working Paper WPS7238, The World Bank, Washington, D.C.

Bruhn, M. and McKenzie, D. (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, **1** (4), 200–232.

Bruns, B., Costa, L. and Cunha, N. (2018). Through the Looking Glass: Can Classroom Observation and Coaching Improve Teacher Performance in Brazil? *Economics of Education Review*, **64**, 214–250.

— and Luque, J. (2014). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Tech. Rep. 89514, The World Bank, Washington, D.C.

Brynjolfsson, E., Rock, D. and Syverson, C. (2017). *Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics*. Tech. Rep. w24001, National Bureau of Economic Research, Cambridge, MA.

Bulman, G. and Fairlie, R. (2016). Technology and Education. In *Handbook of the Economics of Education*, vol. 5, Elsevier, pp. 239–280.

Carrillo, P. E., Onofa, M. and Ponce, J. (2010). *Information Technology and Student Achievement: Evidence from a Randomized Experiment in Ecuador*. Tech. Rep. IDB-WP-223, Inter-American Development Bank, Washington, D.C.

Castro, J. F., Glewwe, P. and Montero, R. (2019). *Work With What You've Got: Improving Teachers' Pedagogical Skills at Scale in Rural Peru*. Working Paper 158, Peruvian Economic Association, Puebla, Mexico.

Cilliers, J., Fleisch, B., Prinsloo, C. and Taylor, S. (2020). How to Improve Teaching Practice? Experimental Comparison of Centralized Training and In-classroom Coaching. *Journal of Human Resources*, **55** (3), 926–962.

Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, Va: Association for Supervision and Curriculum Development, 2nd edn., oCLC: ocm71348683.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K. and Sundararaman, V. (2013). School Inputs, Household Substitution, and Test Scores. *American Economic Journal: Applied Economics*, **5** (2), 29–57.

de la Torre, J. (2009). A Cognitive Diagnosis Model for Cognitively Based Multiple-Choice Options. *Applied Psychological Measurement*, **33** (3), 163–183.

—, Carmona, G., Kieftenbeld, V., Tjoe, H. and Lima, C. (2016). Diagnostic classification models and mathematics education research: Opportunities and challenges. In A. Izsák, J. T. Remillard and J. Templin (eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations*, no. 15 in Journal for Research in Mathematics Education Monograph Series, Reston, VA: National Council of Teachers of Mathematics, pp. 53–71.

de Ree, J., Muralidharan, K., Pradhan, M. and Rogers, H. (2018). Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia. *The Quarterly Journal of Economics*, **133** (2), 993–1039.

deĂlaĂTorre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, **76** (2), 179–199.

Dhar, D., Jain, T. and Jayachandran, S. (2018). *Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India*. Working Paper 25331, National Bureau of Economic Research.

Egert, F., Fukkink, R. G. and Eckhardt, A. G. (2018). Impact of In-Service Professional Development Programs for Early Childhood Teachers on Quality Ratings and Child Outcomes: A Meta-Analysis. *Review of Educational Research*, **88** (3), 401–433.

Escueta, M., Nickow, A. J., Oreopoulos, P. and Quan, V. (2020). Upgrading Education with Technology: Insights from Experimental Research. *Journal of Economic Literature*, **58** (4), 897–996.

Evans, D. K. and Popova, A. (2016). What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews. *The World Bank Research Observer*, **31** (2), 242–270.

Fabregas, R. (2018). Broadcasting Human Capital? The Long Term Effects of Mexico's Telesecundarias.

Ferman, B., Finamor, L. and Lima, L. (2019). *Are Public Schools Ready to Integrate Math Classes with Khan Academy?* Working Paper 94736, University of Munich, Munich.

Fryer, R. G. J. (2017). The Production of Human Capital in Developed Countries. In A. V. Banerjee and E. Duflo (eds.), *Handbook of Economic Field Experiments*, vol. 2, Elsevier, pp. 95–322.

Glewwe, P., Kremer, M. and Moulin, S. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, **1** (1), 112–135.

—, —, — and Zitzewitz, E. (2004). Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya. *Journal of Development Economics*, **74** (1), 251–268.

Graham, C. R. (2006). Blended learning systems: Definition, current trends, and future directions. In C. J. Bonk and C. R. Graham (eds.), *The Handbook of Blended Learning: Global Perspectives, Local Designs*, San Francisco: John Wiley & Sons, pp. 3–21, google-Books-ID: tKdyCwAAQBAJ.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L. and Ball, D. L. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. *Cognition and Instruction*, **26** (4), 430–511.

Jackson, C. K., Rockoff, J. E. and Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. *Annual Review of Economics*, **6** (1), 801–825.

Jacob, B. and Rothstein, J. (2016). The Measurement of Student Ability in Modern Assessment Systems. *Journal of Economic Perspectives*, **30** (3), 85–108.

Jacob, B. A. and Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, **118** (3), 843–877.

Jamison, D. T., Searle, B., Galda, K. and Heyneman, S. P. (1981). Improving elementary mathematics education in Nicaragua: An experimental study of the impact of textbooks and radio on achievement. *Journal of Educational Psychology*, **73** (4), 556–567.

Johnston, J. and Ksoll, C. (2017). *Effectiveness of Interactive Satellite-Transmitted Instruction: Experimental Evidence from Ghanaian Primary Schools*. Working Paper 17-08, Stanford Center for Education Policy Analysis, Stanford, CA.

Kerwin, J. T. and Thornton, R. L. (2021). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. *The Review of Economics and Statistics*, **103** (2), 251–264.

Kolen, M. J. and Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

Kraft, M. A., Blazar, D. and Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*, **88** (4), 547–588.

Lai, F., Luo, R., Zhang, L. and Huang, S., Xinzhe Rozelle (2015). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education*, **47**, 34–48.

Linden, L. L. (2008). *Complement or substitute? The effect of technology on student achievement in India*. Working Paper 17, The World Bank, Washington, D.C.

Loyalka, P., Popova, A., Li, G. and Shi, Z. (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.

Lynch, K., Hill, H. C., Gonzalez, K. E. and Pollard, C. (2019). Strengthening the Research Base That Informs STEM Instructional Improvement Efforts: A Meta-Analysis. *Educational Evaluation and Policy Analysis*, **41** (3), 260–293.

Ma, Y., Fairlie, R., Loyalka, P. and Rozelle, S. (2020). *Isolating the Tech from EdTech: Experimental Evidence on Computer Assisted Learning in China*. Tech. Rep. w26953, National Bureau of Economic Research, Cambridge, MA.

Majerowicz, S. and Montero, R. (2018). Can Teaching be Taught? Experimental Evidence from a Teacher Coaching Program in Peru.

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

Molina, E., Fatima, S. F., Ho, A. D., Melo, C., Wilichowski, T. M. and Pushparatnam, A. (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.

Muralidharan, K., Singh, A. and Ganimian, A. J. (2019). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review*, **109** (4), 1426–1460.

Mutz, D. C., Pemantle, R. and Pham, P. (2018). The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data. *The American Statistician*, pp. 1–11.

Naik, G., Chitre, C., Bhalla, M. and Rajan, J. (2020). Impact of use of technology on student learning outcomes: Evidence from a large-scale experiment in India. *World Development*, **127**, 104736.

Naslund-Hadley, E., Parker, S. W. and Hernandez-Agramonte, J. M. (2014). Fostering Early Math Comprehension: Experimental Evidence from Paraguay. *Global Education Review*, **1** (4), 135–154.

Navarro-Sola, L. (2019). Secondary School Expansion through Televised Lessons: The Labor Market Returns of the Mexican Telesecundaria.

Sabarwal, S., Evans, D. K. and Marshak, A. (2014). *The Permanent Input Hypothesis: The Case of Textbooks and (no) Student Learning in Sierra Leone*. Working Paper 7012, The World Bank, Washington, D.C.

Samejima, F. (1973). A Comment on Birnbaum's Three-Parameter Logistic Model in the Latent Trait Theory. *Psychometrika*, **38** (2), 221–233.

Seo, H. K. (2017). *Do School Electrification and Provision of Digital Media Deliver Educational Benefits? First-year Evidence from 164 Tanzanian Secondary Schools*. Working Paper E-40308-TZA-2, International Growth Centre, London.

Solow, R. M. (1987). We'd better watch out. *New York Times Book Review*.

Stallings, J. A., Knight, S. L. and Markham, D. (2014). *Using the Stallings observation system to investigate time on task in four countries*. Tech. Rep. 92558, The World Bank.

Stocking, M. L. and Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, **7** (2), 201–210.

Taylor, E. S. (2018). New Technology and Teacher Productivity.

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, **18** (6), 3045–3089.

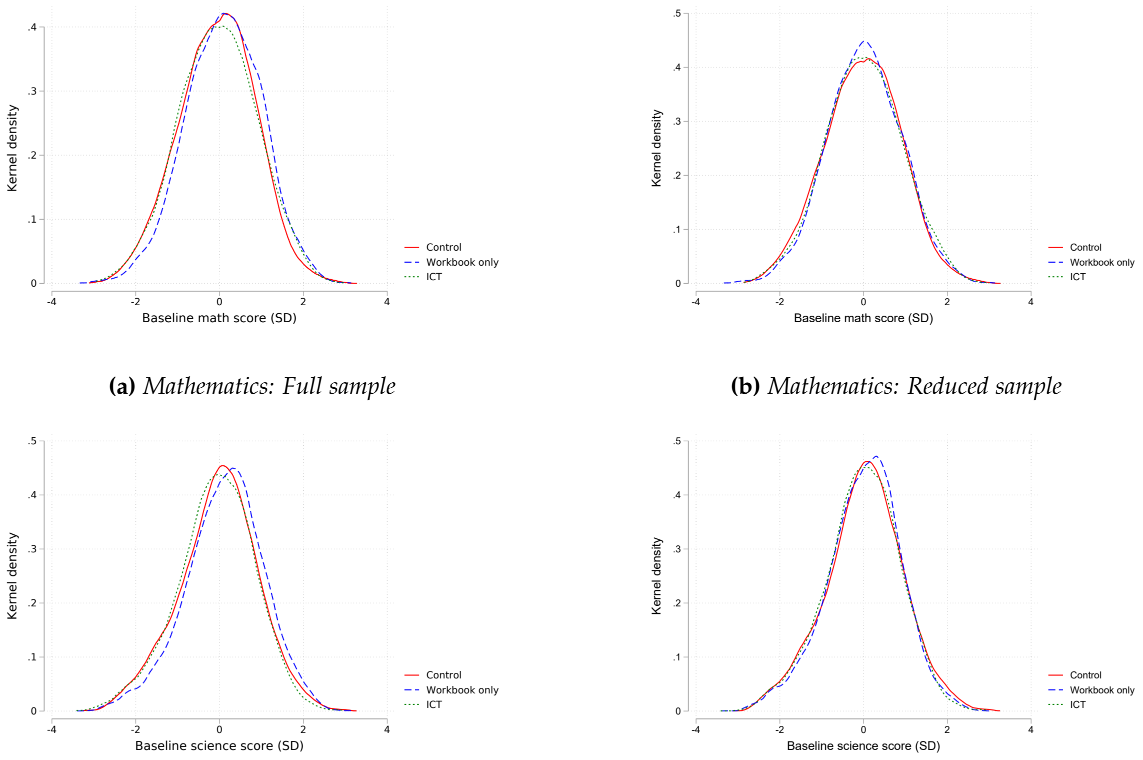
von Hippel, P. T. and Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review*, **64**, 298–312.

Zhang, L., Lai, F., Pang, X., Yi, H. and Rozelle, S. (2013). The Impact of Teacher Training on Teacher and Student Outcomes: Evidence from a Randomised Experiment in Beijing Migrant Schools. *Journal of Development Effectiveness*, **5** (3), 339–358.

Appendices

A Additional figures

Figure A1: *Balance on baseline test scores.*



(a) *Mathematics: Full sample*

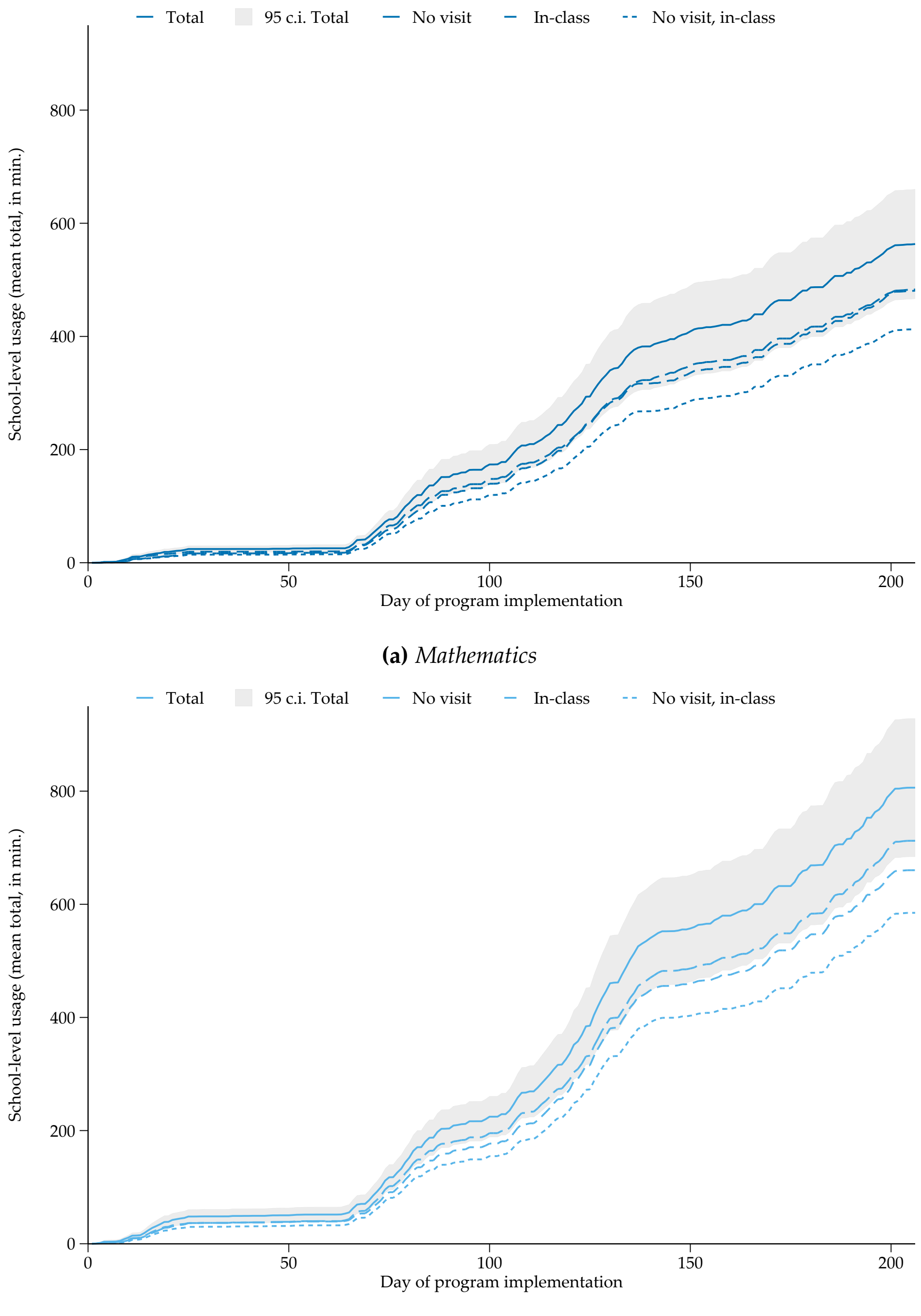
(b) *Mathematics: Reduced sample*

(c) *Science: Full sample*

(d) *Science: Reduced sample*

Note: This figure reports on the sample's balance across the three groups, as per the baseline tests in mathematics and science. Each panel shows kernel density plots, by treatment status, of residuals from a regression of baseline test scores on strata fixed effects. The top panels report results for mathematics; the bottom panels report results for science. Left panels show the full sample; the ICT and control groups are balanced, but students in the workbook group systematically outperform students in the other two groups. Right panels show a reduced sample of schools, where (the 21 schools of) the seven most severely imbalanced randomization triplets are dropped.

Figure A2: *Cumulative software usage in ICT schools, over time*

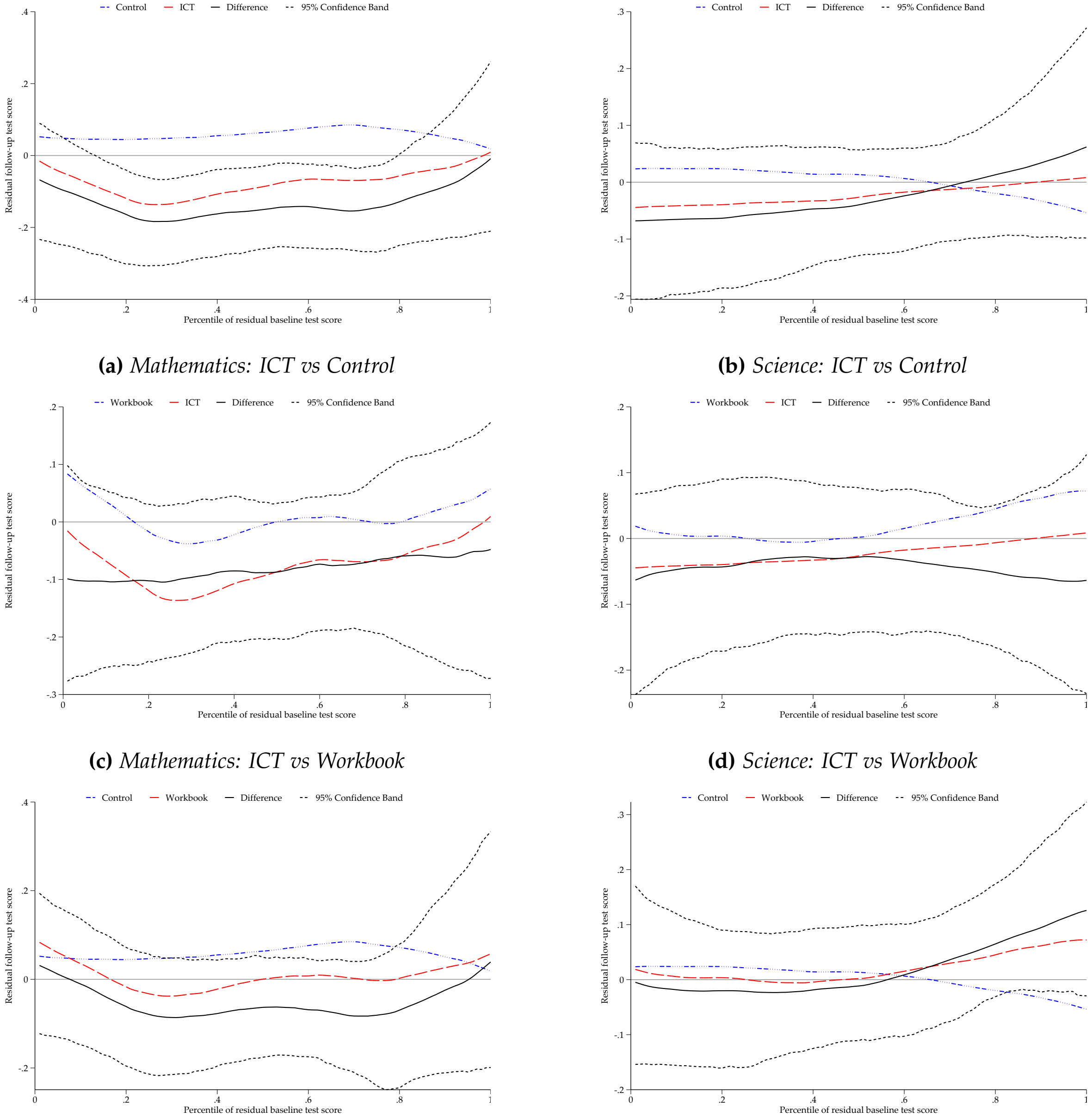


(a) *Mathematics*

(b) *Science*

Notes. By subject, these figures present the mean, school-level, cumulative usage of Avanti's software (in minutes), for the 80 ICT schools. They cover the period from May 7 to November 28, 2019 (inclusive). 95 percent confidence intervals for total usage in grey. "No visit" refers to usage on a day without a school visit from the NGO. "In-class" refers to usage of the software version intended for in-class usage. "No visit, in-class" refers to usage of the software version intended for in-class usage, on a day without a school visit from the NGO.

Figure A3: *Non-parametric investigation of ITT effects by percentiles of baseline scores*



(a) *Mathematics: ICT vs Control*

(b) *Science: ICT vs Control*

(c) *Mathematics: ICT vs Workbook*

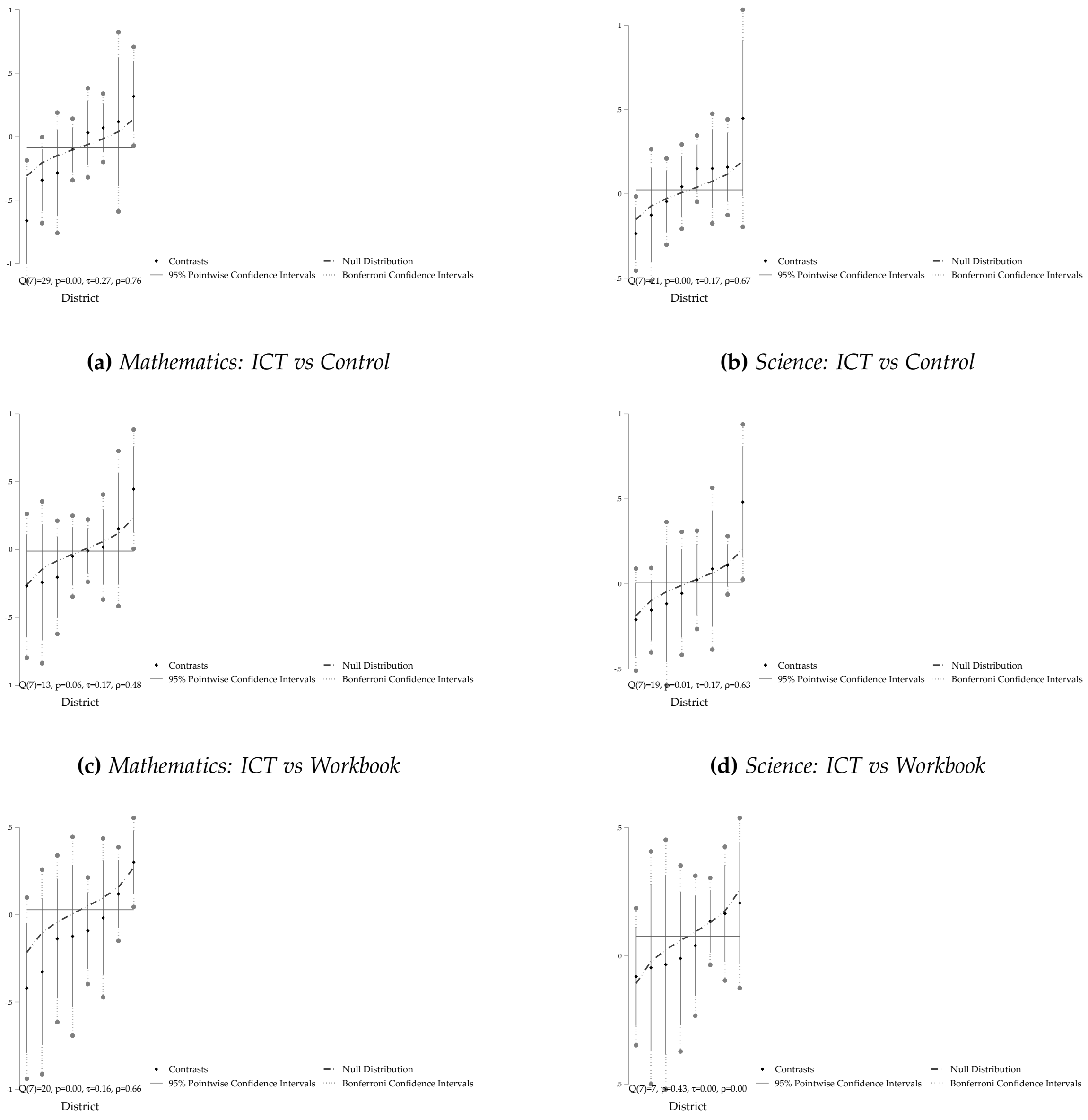
(d) *Science: ICT vs Workbook*

(e) *Mathematics: Workbook vs Control*

(f) *Science: Workbook vs Control*

Notes. These figures provide a non-parametric investigation of ITT effects by percentiles of baseline scores. The treatment and control lines are estimated using local linear regressions. The pointwise treatment effects are calculated as the difference. The 95% confidence intervals are estimated using bootstrapping; bootstrap iterations are blocked at the school-level, to allow for the clustering of standard errors. The x-axis is the percentile of a students test score at baseline. The y-axis is the residual of a regression of a student's test score at follow-up on randomization strata fixed effects and a vector of student-, school-, and village-level covariates, selected via LASSO (see Appendix E). "Baseline" refers to the assessment conducted in December 2018. "Follow-up" refers to the assessment conducted in November 2019.

Figure A4: *Heterogeneity in ITT effects across districts*



(a) *Mathematics: ICT vs Control*

(b) *Science: ICT vs Control*

(c) *Mathematics: ICT vs Workbook*

(d) *Science: ICT vs Workbook*

(e) *Mathematics: Workbook vs Control*

(f) *Science: Workbook vs Control*

Notes. These figures provide "caterpillar plots" of ITT effects by district (cf. von Hippel and Bellows, 2018). Each black dot refers to the point estimate for a given district. All estimations include randomization strata fixed effects (F.E.s) and a vector of school- and village-level covariates, selected via LASSO (see Appendix E). Confidence intervals allow for clustering of standard errors at the school level. Bonferroni confidence intervals adjust standard errors for multiple hypothesis testing. The black solid line shows the null distribution of "effects" that can be expected due to error. τ is the heterogeneity standard deviation. Q refers to Cochran's Q statistic, which follows a χ^2 distribution, and p reports on the corresponding p-value for a test of the null hypothesis of no heterogeneity. ρ estimates the reliability; that is, the share of variance in estimates that is attributable to heterogeneity (rather than error).

B Additional tables

Table A1: *Program components, partner responsibilities, and intervention groups*

	Avanti Role	GoH / HSSPP Role	80 GSSS randomly assigned to **ICT Group** (Group 1)	80 GSSS randomly assigned to **Workbook Group** (Group 2)	80 GSSS randomly assigned to **Control Group** (Group 3)
Capacity building workshops for teachers	Training design and implementation	Provision of master trainers from the HSSP	Provided to all teachers	Provided to all teachers	No training
ICT infrastructure, digital learning materials for blended instruction	NIL	HSSPP to purchase install, and maintain projector/televisions, computers and sound systems	2 Smart Classrooms set up per school, digital materials	No ICT infrastructure, no digital materials	No ICT infrastructure, no digital materials
Workbooks	Avanti to design and provide pdfs for printing	HSSPP to print and distribute workbooks	Workbooks provided to all students	Workbooks provided to all students	No Workbooks provided
Assessments	Design and provision of test papers and OMR pdfs to HSSPP for printing. Support in invigilation and spot checks	HSSPP to print, distribute and conduct the test through BRPs/ABRCs	Baseline, midline, endline tests conducted	Baseline, midline, endline tests conducted	Baseline, midline, endline tests conducted
Classroom observations	Observations by Avanti Program Managers	Observations by Master Trainers	Monthly observation (4 classrooms per visit)	Monthly observation (4 classrooms per visit)	Six-weekly observation (4 classrooms per visit)

Notes. ICT: Information and communication technology. GoH: Government of Haryana. DIET: District Institute of Education and Training. HSSPP: Haryana School Shiksha Pariyojna Parishad. GSSS: Government Senior Secondary School.

Table A2: *Summary measures of ICT infrastructure and program availability*

	Number of observations			Mean			Differences (F.E.s)		
	Control	ICT	Workbook	Control	ICT	Workbook	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Rooms available at baseline									
One functioning smart classrooms or more	80	80	80	86.25	82.50	86.25	-3.75	-3.75	-0.00
				[34.65]	[38.24]	[34.65]	(5.33)	(5.33)	(5.33)
Two functioning smart classrooms or more	80	80	80	22.50	22.50	18.75	-0.00	3.75	-3.75
				[42.02]	[42.02]	[39.28]	(6.07)	(6.07)	(6.07)
Panel B: Infrastructure exists at follow-up									
Generator	48	80	69	56.25	81.25	63.77	25.45***	20.63***	4.81
				[50.13]	[39.28]	[48.42]	(7.87)	(6.80)	(8.15)
Smart TV	48	80	69	35.42	100.00	43.48	65.34***	56.58***	8.76
				[48.33]	[.]	[49.94]	(6.93)	(5.99)	(7.17)
Speaker	48	80	69	43.75	91.25	52.17	47.94***	40.67***	7.27
				[50.13]	[28.43]	[50.32]	(7.60)	(6.57)	(7.87)
Tablet	48	80	69	8.33	100.00	13.04	92.11***	87.30***	4.81
				[27.93]	[.]	[33.92]	(4.78)	(4.14)	(4.95)
Panel C: Infrastructure functional at follow-up									
Electricity	48	80	69	43.75	96.25	42.03	51.83***	56.44***	-4.61
				[50.13]	[19.12]	[49.72]	(8.13)	(7.03)	(8.42)
Generator	48	80	69	39.58	72.50	31.88	31.50***	43.56***	-12.06
				[49.42]	[44.93]	[46.94]	(8.89)	(7.69)	(9.21)
Smart TV	48	80	69	29.17	98.75	28.99	69.85***	70.15***	-0.30
				[45.93]	[11.18]	[45.70]	(7.12)	(6.16)	(7.38)
Speaker	48	80	69	37.50	90.00	46.38	52.82***	45.53***	7.29
				[48.92]	[30.19]	[50.23]	(7.93)	(6.85)	(8.21)
Tablet	48	80	69	8.33	98.75	8.70	91.08***	90.97***	0.11
				[27.93]	[11.18]	[28.38]	(4.45)	(3.84)	(4.60)
Panel D: Any ICT program at follow-up									
Any ICT program active	48	80	69	4.17	100.00	2.90	96.81***	96.52***	0.30
				[20.19]	[.]	[16.90]	(2.85)	(2.46)	(2.95)

Notes. This table provides descriptive statistics for the study sample, by treatment status, on school-level measures of ICT infrastructure. "Baseline" refers to an infrastructure survey, conducted in December 2018. "Follow-up" refers to data from the most recent school visit, conducted in October-December 2019. Not all Control and Workbook schools have been surveyed yet, but the order of school visits is random, and the results are therefore representative. Standard deviations in brackets; standard errors in parentheses. All estimations include randomization strata fixed effects (F.E.s). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A3: *Sensitivity of results for main learning outcomes*

	Follow-up differences (F.E.s + Controls)		
	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)
Panel A: Randomization inference			
Mathematics	-0.16	-0.09	-0.07
	[0.04]	[0.23]	[0.36]
Science	-0.04	-0.05	0.02
	[0.52]	[0.33]	[0.75]
Panel B: Reduced sample			
Mathematics	-0.13	-0.06	-0.07
	(0.06)*	(0.06)	(0.07)
Science	-0.01	-0.03	0.03
	(0.05)	(0.05)	(0.05)

Notes. This table investigates the sensitivity of results for the study's main results (compare to Table 4, Panel A). Panel A uses randomization inference, replicating the study's randomization strategy in each iteration. Panel B removes imbalanced randomization strata. "Follow-up" refers to the assessment conducted in November 2019. All estimations include randomization strata fixed effects (F.E.s). "Controls" indicates the inclusion of a vector of student-, school-, and village-level covariates, selected via LASSO (see Appendix E). P-values in brackes; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%. *Sample.* 18,615 grade-9 and grade-10 students, in the study's 240 schools.

Table A4: *ITT effects on instructional quality and teaching practices (mathematics)*

	Follow-up (most recent school visit)			Follow-up differences (F.E.s + Controls)			Follow-up differences (F.E.s + Controls + Adjust.)		
	Control	ICT	Workbook	ICT vs Control	ICT vs Workbook	Workbook vs Control	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Effects on observed instructional quality									
Monitoring of student learning	0.12	-0.14	0.07	-0.31**	-0.33**	0.02	-0.89***	-0.48***	-0.41***
	[0.98]	[1.01]	[1.09]	(0.15)	(0.15)	(0.14)	(0.15)	(0.14)	(0.12)
Feedback	0.04	-0.17	0.09	-0.19	-0.35***	0.16	-0.63***	-0.54***	-0.09
	[0.94]	[0.92]	[1.15]	(0.12)	(0.14)	(0.14)	(0.11)	(0.12)	(0.13)
Management of class time	0.15	-0.28	-0.14	-0.60***	-0.26	-0.34**	-1.31***	-0.60***	-0.71***
	[1.01]	[1.32]	[1.05]	(0.16)	(0.17)	(0.16)	(0.16)	(0.17)	(0.16)
Dense focus on math/science	0.08	-0.41	-0.19	-0.51***	-0.25	-0.26*	-1.57***	-0.93***	-0.64***
	[1.05]	[1.30]	[1.11]	(0.17)	(0.17)	(0.15)	(0.17)	(0.17)	(0.15)
Clarity, lack of errors	0.16	-0.06	-0.10	-0.44***	-0.16	-0.29*	-0.71***	-0.72***	0.01
	[0.94]	[1.17]	[1.11]	(0.14)	(0.16)	(0.15)	(0.14)	(0.16)	(0.15)
Richness	-0.06	-0.27	-0.23	-0.27**	-0.00	-0.27**	-1.58***	-0.92***	-0.67***
	[0.99]	[0.89]	[0.93]	(0.13)	(0.14)	(0.13)	(0.11)	(0.12)	(0.11)
QUIP (Index)	0.11	-0.21	-0.07	-0.46***	-0.22	-0.24	-1.42***	-0.89***	-0.52***
	[0.98]	[0.99]	[1.10]	(0.15)	(0.16)	(0.16)	(0.14)	(0.15)	(0.15)
Panel B: Effects on observed teaching practices									
Instruction (% of class time)	0.67	0.57	0.59	-0.14***	-0.05	-0.09**	-0.13***	-0.04	-0.09**
	[0.29]	[0.32]	[0.33]	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Management (% of class time)	0.08	0.09	0.08	0.02	0.02	-0.00	0.02	0.02	-0.00
	[0.14]	[0.13]	[0.12]	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Off-task (% of class time)	0.26	0.34	0.33	0.12***	0.03	0.10**	0.11***	0.02	0.09**
	[0.29]	[0.36]	[0.34]	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Class held in smart classroom (% of classes)	0.00	0.65	0.00	0.67***	0.63***	0.03	0.66***	0.63***	0.04
	[.]	[0.48]	[.]	(0.04)	(0.04)	(0.02)	(0.05)	(0.04)	(0.03)
Use of ICT (% of classes)	0.00	0.56	0.00	0.54***	0.51***	0.03	0.52***	0.50***	0.03
	[.]	[0.50]	[.]	(0.05)	(0.05)	(0.02)	(0.05)	(0.05)	(0.03)
Use of ICT (% of class time)	0.00	0.32	0.00	0.29***	0.27***	0.02	0.28***	0.26***	0.02
	[.]	[0.34]	[.]	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
Use of textbooks (% of classes)	0.39	0.14	0.41	-0.24***	-0.26***	0.02	-0.21***	-0.25***	0.04
	[0.49]	[0.35]	[0.49]	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)	(0.07)
Use of textbooks (% of class time)	0.14	0.06	0.17	-0.08***	-0.11***	0.03	-0.08***	-0.11***	0.03
	[0.23]	[0.18]	[0.26]	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Use of notebooks (% of classes)	0.25	0.15	0.29	-0.11**	-0.11*	0.00	-0.11**	-0.10**	-0.01
	[0.44]	[0.36]	[0.46]	(0.05)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)
Use of notebooks (% of class time)	0.09	0.06	0.11	-0.04	-0.03	-0.01	-0.05**	-0.03	-0.01
	[0.20]	[0.16]	[0.23]	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Group activity (% of classes)	0.01	0.00	0.00	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01
	[0.09]	[.]	[.]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Group activity (% of class time)	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00
	[0.02]	[.]	[.]	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Panel C: Effects on student-reported teaching practices									
Student easily understands what the teacher teaches	-0.00	0.06	-0.04	0.19	0.11	0.09	0.25**	0.10	0.15
	[0.99]	[0.97]	[1.00]	(0.12)	(0.11)	(0.12)	(0.12)	(0.10)	(0.12)
Teacher gives interesting things to do in class	0.03	0.11	-0.10	0.05	0.22*	-0.17	0.09	0.16	-0.06
	[0.99]	[0.93]	[1.03]	(0.15)	(0.13)	(0.13)	(0.14)	(0.12)	(0.13)
Teacher explains topic again if students do not understand	0.03	-0.08	0.09	-0.00	-0.08	0.07	-0.07	-0.08	0.01
	[0.89]	[1.15]	[0.90]	(0.11)	(0.10)	(0.12)	(0.11)	(0.10)	(0.13)
Teacher does a variety of things to help learn	0.10	0.06	0.10	0.01	0.07	-0.06	0.03	0.08	-0.05
	[0.93]	[1.03]	[0.94]	(0.14)	(0.12)	(0.13)	(0.14)	(0.11)	(0.12)
Index	0.04	-0.00	0.04	0.07	0.04	0.04	0.06	0.02	0.04
	[0.96]	[1.08]	[0.99]	(0.12)	(0.11)	(0.12)	(0.12)	(0.11)	(0.13)
Teacher used videos to teach, past week	0.03	0.59	0.01	0.59***	0.58***	0.02	0.60***	0.58***	0.03
	[0.16]	[0.49]	[0.08]	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Student usally works with at least one peer	0.86	0.86	0.80	0.03	0.05	-0.03	0.02	0.06*	-0.04
	[0.34]	[0.35]	[0.40]	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
Student usually works in groups	0.45	0.41	0.37	-0.12	0.01	-0.14*	-0.04	-0.01	-0.03
	[0.50]	[0.49]	[0.49]	(0.08)	(0.07)	(0.08)	(0.07)	(0.05)	(0.08)
# of math / science classes, past week	5.39	5.06	4.87	-0.32	0.21	-0.53	-0.30	0.28	-0.58**
	[1.71]	[1.89]	[1.83]	(0.31)	(0.27)	(0.32)	(0.24)	(0.18)	(0.24)

Notes. This table presents the intent-to-treat (ITT) effects of the interventions on lesson-level measures of instructional quality and teaching practices. "Index" refers to the inverse covariance matrix-weighted average, following Anderson (2008). "Follow-up" refers to data from the most recent school visit, conducted in October-November 2019. All estimations include randomization strata fixed effects (F.E.s). All estimations include all observations (not just those from the most recent school visit) and interact treatment effects with a linear time trend; the reference date is the most recent observation date (November 15, 2019). "Controls" indicates the inclusion of a vector of school- and village-level covariates, selected via LASSO (see Appendix E). In Panel A, "adjust." refers to an adjustment for systematic differences in comparison to video-based re-ratings (see Appendix C), and the inclusion of rater fixed effects; in Panels B and C, "adjust." refers to the inclusion of rater fixed effects. Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.

Sample. Panel A and Panel B from 693 classroom observations in mathematics. Panel C from 604 student interviews about mathematics. School visits follow a random schedule, classrooms and students are subsampled at random, and interview questions about subjects (mathematics vs science) are assigned at random.

Table A5: *ITT effects on instructional quality and teaching practices (science)*

	Follow-up (most recent school visit)			Follow-up differences (F.E.s + Controls)			Follow-up differences (F.E.s + Controls + Adjust.)		
	Control	ICT	Workbook	ICT vs Control	ICT vs Workbook	Workbook vs Control	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Effects on observed instructional quality									
Monitoring of student learning	-0.07	-0.34	-0.01	-0.34**	-0.46***	0.13	-0.67***	-0.50***	-0.17
	[0.99]	[1.06]	[1.03]	(0.13)	(0.12)	(0.14)	(0.13)	(0.13)	(0.13)
Feedback	0.00	-0.22	0.16	-0.27**	-0.40***	0.14	-0.22*	-0.72***	0.50***
	[1.03]	[0.93]	[1.12]	(0.13)	(0.15)	(0.15)	(0.12)	(0.13)	(0.13)
Management of class time	0.15	-0.14	-0.22	-0.28*	0.06	-0.35**	-0.72***	-0.43***	-0.29*
	[0.83]	[1.15]	[1.20]	(0.15)	(0.16)	(0.15)	(0.15)	(0.16)	(0.15)
Dense focus on math/science	0.07	-0.26	-0.35	-0.30**	0.06	-0.36***	-0.40***	-0.95***	0.55***
	[0.93]	[1.14]	[1.26]	(0.15)	(0.15)	(0.13)	(0.14)	(0.14)	(0.12)
Clarity, lack of errors	-0.03	-0.13	-0.11	-0.10	-0.10	-0.01	-0.14	-0.58***	0.44***
	[1.07]	[1.19]	[1.20]	(0.17)	(0.17)	(0.16)	(0.17)	(0.17)	(0.16)
Richness	0.04	-0.19	-0.12	-0.19	-0.08	-0.10	-1.47***	-1.29***	-0.18
	[0.99]	[0.87]	[0.92]	(0.14)	(0.14)	(0.13)	(0.14)	(0.13)	(0.11)
QUIP (Index)	0.06	-0.31	-0.07	-0.40***	-0.36**	-0.04	-0.83***	-1.19***	0.36***
	[1.00]	[1.11]	[1.20]	(0.15)	(0.15)	(0.15)	(0.14)	(0.15)	(0.13)
Panel B: Effects on observed teaching practices									
Instruction (% of class time)	0.65	0.56	0.61	-0.10**	-0.04	-0.06	-0.07*	-0.02	-0.05
	[0.32]	[0.32]	[0.33]	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Management (% of class time)	0.04	0.12	0.07	0.08***	0.05***	0.03**	0.08***	0.06***	0.03**
	[0.08]	[0.14]	[0.10]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Off-task (% of class time)	0.31	0.32	0.33	0.02	-0.02	0.04	-0.01	-0.04	0.03
	[0.33]	[0.35]	[0.35]	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Class held in smart classroom (% of classes)	0.00	0.66	0.00	0.65***	0.66***	-0.01	0.64***	0.65***	-0.01
	[.]	[0.48]	[.]	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)
Use of ICT (% of classes)	0.00	0.59	0.00	0.58***	0.59***	-0.01	0.58***	0.59***	-0.01
	[.]	[0.49]	[.]	(0.04)	(0.04)	(0.03)	(0.05)	(0.04)	(0.03)
Use of ICT (% of class time)	0.00	0.36	0.00	0.34***	0.35***	-0.01	0.34***	0.34***	-0.00
	[.]	[0.34]	[.]	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
Use of textbooks (% of classes)	0.43	0.13	0.56	-0.24***	-0.34***	0.10*	-0.26***	-0.36***	0.11*
	[0.50]	[0.34]	[0.50]	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Use of textbooks (% of class time)	0.19	0.04	0.24	-0.12***	-0.18***	0.05*	-0.14***	-0.19***	0.05
	[0.27]	[0.12]	[0.27]	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Use of notebooks (% of classes)	0.12	0.10	0.28	0.03	-0.13***	0.15***	0.02	-0.15***	0.17***
	[0.32]	[0.31]	[0.45]	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Use of notebooks (% of class time)	0.04	0.03	0.08	-0.00	-0.05***	0.05**	-0.01	-0.06***	0.05**
	[0.13]	[0.12]	[0.17]	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Group activity (% of classes)	0.01	0.01	0.03	0.00	-0.01	0.02	0.00	-0.01	0.02
	[0.10]	[0.10]	[0.16]	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)
Group activity (% of class time)	0.00	0.00	0.01	0.00	-0.01	0.01	0.00	-0.01	0.01
	[0.02]	[0.02]	[0.05]	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Panel C: Effects on student-reported teaching practices									
Student easily understands what the teacher teaches	-0.07	0.11	0.09	-0.04	0.08	-0.12	-0.05	0.05	-0.11
	[1.05]	[0.93]	[0.90]	(0.13)	(0.10)	(0.12)	(0.14)	(0.10)	(0.12)
Teacher gives interesting things to do in class	0.07	0.04	-0.03	-0.31**	-0.08	-0.23*	-0.24*	-0.09	-0.15
	[0.98]	[1.01]	[1.09]	(0.15)	(0.12)	(0.13)	(0.15)	(0.12)	(0.13)
Teacher explains topic again if students do not understand	0.04	-0.05	0.11	-0.09	-0.18	0.08	-0.15	-0.18	0.03
	[0.95]	[1.22]	[0.78]	(0.14)	(0.12)	(0.15)	(0.14)	(0.12)	(0.15)
Teacher does a variety of things to help learn	-0.01	0.10	-0.02	0.11	0.15	-0.05	0.10	0.22**	-0.12
	[1.05]	[0.93]	[0.98]	(0.12)	(0.11)	(0.13)	(0.11)	(0.11)	(0.12)
Index	0.01	0.05	0.10	-0.08	-0.04	-0.04	-0.12	-0.03	-0.09
	[0.99]	[1.02]	[0.91]	(0.15)	(0.12)	(0.16)	(0.15)	(0.12)	(0.16)
Teacher used videos to teach, past week	0.03	0.57	0.01	0.59***	0.60***	-0.01	0.59***	0.60***	-0.01
	[0.16]	[0.50]	[0.11]	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Student usally works with at least one peer	0.88	0.80	0.78	-0.06	0.04	-0.10**	-0.04	0.06	-0.10**
	[0.32]	[0.40]	[0.41]	(0.05)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)
Student usually works in groups	0.48	0.38	0.38	-0.17**	-0.02	-0.15**	-0.08	0.01	-0.08
	[0.50]	[0.49]	[0.49]	(0.07)	(0.06)	(0.07)	(0.06)	(0.05)	(0.06)
# of math / science classes, past week	5.34	5.03	4.98	-0.25	0.16	-0.41	-0.09	0.30	-0.39
	[1.63]	[1.85]	[1.83]	(0.30)	(0.27)	(0.31)	(0.24)	(0.21)	(0.26)

Notes. This table presents the intent-to-treat (ITT) effects of the interventions on lesson-level measures of instructional quality and teaching practices. "Index" refers to the inverse covariance matrix-weighted average, following Anderson (2008). "Follow-up" refers to data from the most recent school visit, conducted in October-November 2019. All estimations include randomization strata fixed effects (F.E.s). All estimations include all observations (not just those from the most recent school visit) and interact treatment effects with a linear time trend; the reference date is the most recent observation date (November 15, 2019). "Controls" indicates the inclusion of a vector of school- and village-level covariates, selected via LASSO (see Appendix E). In Panel A, "adjust." refers to an adjustment for systematic differences in comparison to video-based re-ratings (see Appendix C), and the inclusion of rater fixed effects; in Panels B and C, "adjust." refers to the inclusion of rater fixed effects. Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.

Sample. Panel A and Panel B from 650 classroom observations in science. Panel C from 610 student interviews about science. School visits follow a random schedule, classrooms and students are subsampled at random, and interview questions about subjects (mathematics vs science) are assigned at random.

Table A6: *ITT effects on student perceptions and attitudes towards mathematics and science (by subject)*

	Follow-up (most recent school visit)			Follow-up differences (F.E.s + Controls)			Follow-up differences (F.E.s + Controls + Adjust.)		
	Control	ICT	Workbook	ICT vs Control	ICT vs Workbook	Workbook vs Control	ICT vs Control	ICT vs Workbook	Workbook vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Mathematics									
I enjoy learning mathematics	0.03	-0.01	0.02	-0.25**	0.01	-0.26**	-0.31***	0.01	-0.32**
	[1.03]	[1.08]	[0.99]	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)	(0.13)
I learn many interesting things in mathematics	0.03	-0.09	-0.06	-0.26**	0.00	-0.26**	-0.32***	0.02	-0.34***
	[1.03]	[0.94]	[1.02]	(0.12)	(0.12)	(0.13)	(0.11)	(0.11)	(0.12)
Mathematics makes me nervous (reversed)	-0.06	-0.04	-0.03	-0.24*	-0.08	-0.16	-0.33***	-0.04	-0.28**
	[1.02]	[1.01]	[0.97]	(0.13)	(0.12)	(0.13)	(0.11)	(0.11)	(0.12)
Mathematics is harder than other subjects (reversed)	-0.01	-0.08	-0.04	-0.27**	-0.03	-0.24**	-0.31***	-0.03	-0.28***
	[1.02]	[1.02]	[1.02]	(0.11)	(0.10)	(0.11)	(0.10)	(0.10)	(0.11)
I don't understand what is taught in mathematics (reversed)	0.05	0.02	0.11	-0.17	-0.16	-0.01	-0.23*	-0.17	-0.06
	[1.02]	[1.01]	[1.10]	(0.14)	(0.12)	(0.13)	(0.13)	(0.12)	(0.13)
Index	0.02	-0.06	0.00	-0.35***	-0.08	-0.27**	-0.44***	-0.07	-0.37***
	[1.03]	[1.05]	[1.08]	(0.13)	(0.13)	(0.13)	(0.11)	(0.12)	(0.12)
Panel B: Science									
I enjoy learning science	0.03	-0.17	-0.10	-0.08	-0.05	-0.03	-0.07	-0.02	-0.05
	[1.07]	[0.72]	[0.87]	(0.12)	(0.09)	(0.12)	(0.12)	(0.09)	(0.12)
I learn many interesting things in science	-0.02	-0.01	0.06	0.07	-0.11	0.18	0.07	-0.12	0.19
	[0.98]	[0.96]	[0.95]	(0.13)	(0.11)	(0.12)	(0.12)	(0.12)	(0.12)
Science makes me nervous (reversed)	0.03	-0.05	-0.12	-0.15	0.00	-0.15	-0.14	0.01	-0.14
	[1.00]	[0.90]	[0.90]	(0.12)	(0.10)	(0.12)	(0.12)	(0.10)	(0.12)
Science is harder than other subjects (reversed)	0.08	0.06	-0.04	-0.06	0.06	-0.12	-0.05	0.13	-0.19*
	[0.99]	[0.98]	[0.96]	(0.13)	(0.11)	(0.13)	(0.12)	(0.10)	(0.11)
I don't understand what is taught in science (reversed)	0.08	0.08	0.05	-0.05	0.04	-0.09	-0.06	0.06	-0.12
	[1.01]	[1.07]	[1.09]	(0.13)	(0.12)	(0.13)	(0.13)	(0.11)	(0.13)
Index	0.06	-0.04	-0.05	-0.08	-0.03	-0.06	-0.08	0.01	-0.09
	[1.01]	[0.94]	[0.98]	(0.13)	(0.11)	(0.13)	(0.12)	(0.11)	(0.12)

Notes. This table presents the intent-to-treat (ITT) effects of the interventions on student-level perceptions and attitudes towards mathematics and science. "Index" refers to the inverse covariance matrix-weighted average of the five questions, following Anderson (2008). "Follow-up" refers to data from the most recent school visit, conducted in October-November 2019. All estimations include randomization strata fixed effects (F.E.s). All estimations include all observations (not just those from the most recent school visit) and interact treatment effects with a linear time trend; the reference date is the most recent observation date (November 15, 2019). "Controls" indicates the inclusion of a vector of student, school-, and village-level covariates, selected via LASSO (see Appendix E). "Adjust." refers to the inclusion of interviewer fixed effects. Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.

Sample. 1,214 student interviews (604 about mathematics and 610 about science). School visits follow a random schedule, students are subsampled at random, and interview questions about subjects (mathematics vs science) are assigned at random.

C Measuring instructional quality

C.1 Objectives

The study administered classroom observations to measure the quality of instruction students receive, in mathematics an science. In doing so, I followed Ho's validation framework (cf. Molina *et al.*, 2020, 4). Accordingly, I aimed for theoretical and practical relevance of the measure's content, for the measure to be internally consistent and precise, and for accurate interpretation of the measure across its raters. I also investigated whether the measure is predictive of student learning.

C.2 Content

C.2.1 Domains

The QUIP instrument focuses on six aspects of instructional quality, which are aligned with the program's Theory of Change. The instrument taps into six elements, which are grouped into three pairs: Monitoring of student learning; Feedback (Pair A); Maximization of learning time; Classroom work is mathematically / scientifically dense (Pair B); Presentation of content is clear and not distorted; Richness of mathematics / science (Pair C). Each of the six dimensions is further divided into three sub-dimensions.

The instrument builds on other, well-validated classroom observation instruments – such as MQI (Hill *et al.*, 2008), the Danielson Framework (Danielson, 2007), and CLASS (Allen *et al.*, 2011). Over a one-year process prior to the baseline assessment, I adapted the instrument to the local context and piloted it (out-of-sample), in collaboration with staff at J-PAL South Asia and Avanti Fellows.

C.2.2 Rating categories

Ratings are given on a four-point scale with the following categories: "newcomer", "basic", "proficient", and "exemplary". The 18 sub-dimensions are clearly defined and accompanied by vignettes that clarify the four rating categories, separately, for each sub-dimension. If a dimension was not used at all (e.g., if a teacher did not provide any feedback), observers rated the given dimension as "newcomer". A detailed observer handbook with these definitions is available upon request.

C.3 Administration

C.3.1 Field operations

To guarantee representativeness, each observer followed a pre-determined, random order of school visits. By design, the ratio of school visits between ICT schools, and Workbook schools, and Control schools is 3:3:2. To allow for logistical flexibility, observers could freely schedule school visits, but they could not skip more than five schools in their assigned roster of visits.

During each school visit, four classrooms were randomly selected for observations. My protocol selects one grade-9 and one grade-10 classroom, in mathematics and science, respectively. Students were observed independently of where their class takes place, and independently of who teaches the class.

Per lesson, observers rated four snippets of approximately six minutes each, following prompts on handheld tablets.[20] To reduce the cognitive burden for observers and to increase the quality of ratings, each snippet required the observation and rating of four dimensions (instead of six). Each snippet randomly prompted the rating of two pairs of dimensions (AB, AC, or BC), and I implemented a constraint such that each dimension was rated at least twice per lesson. There are 36 possible group-pair by snippet combinations, and I assigned these combinations to observations at random.

C.3.2 Quality control

Data collection was subjected to three types of quality control mechanisms, as follows. First, incoming data was analyzed on a weekly basis, to reveal inconsistent data points ("high-frequency checks"). Second, a randomly assigned 20 percent of classroom observations were jointly conducted by the observer and her supervisor—during these visits, the supervisor rated a randomly selected half of the time snippets.[21] Third, during these supervisor accompaniments, the supervisor video-recorded the other half of the time snippets—they were then centrally re-rated, twice, by an external team of trained raters.

C.4 Scoring

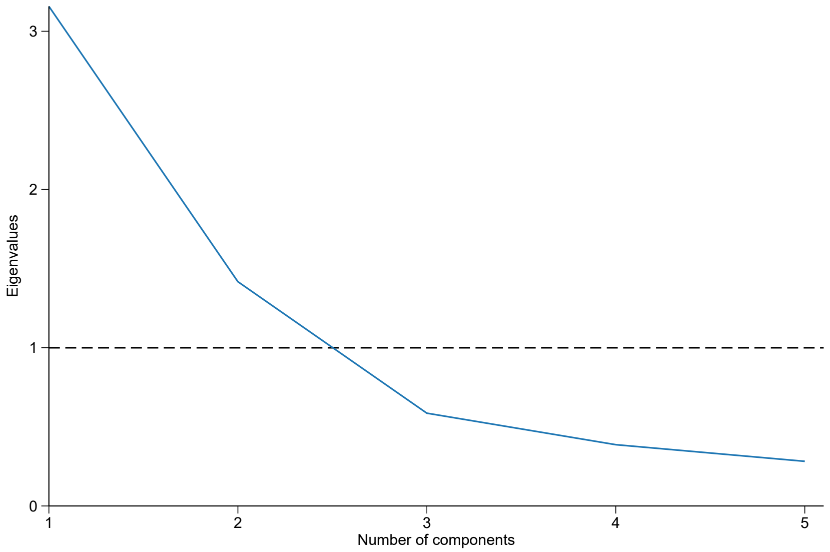
C.4.1 Preferred approach

For each observed lesson, for each dimension, I retain the best rating. Thus, I take into account that some instructional dimensions may not have been used during a given snippet. Thereafter, I standardize the ratings for each dimension, using the control group as reference (mean zero, standard deviation of one). I also construct a summary index across the six dimensions, by calculating their inverse covariance-matrix-weighted average (following Anderson, 2008). In doing so, I recognize that instructional quality may not be unidimensional, and I give greater weight to those dimensions that do not correlate well with others.

[20] Prompts for ratings for the "Stallings" instrument (on instructional practices), and the entry of additional information (e.g., on implementation fidelity) occurred at other times, separately from the QUIP instrument.

[21] I synced the tablets across observers and supervisors, such that both rated the same, randomly selected dimensions, during each snippet. However, they were asked to sit separately and to provide independent ratings.

Figure A5: *QUIP dimensionality*



Notes. This figure provides a Scree plot of eigenvalues, showing the variation accounted for by each element (out of six). This is estimated from a 1-factor polychoric exploratory factor analysis.

C.4.2 Alternative approach, dimensionality

I briefly explored, but did not use two alternative approaches to creating an index measure of instructional quality: factor analysis, and item-response theory (cf. Molina *et al.*, 2020). As discussed below, both approaches do not result as appropriate strategies to create a single index, as they down-weight dimensions of instructional quality that do not correlate well with others.

In Figure A5, I show a scree plot from a polychoric exploratory factor analysis. The analysis suggests that the six dimensions relate to at least two broader aspects of instructional quality. Extracting only the first factor would therefore down-weight information of the second.

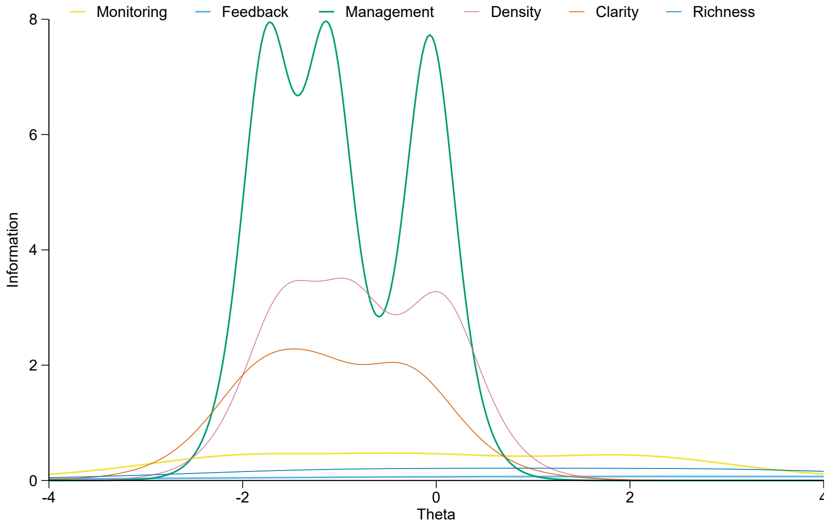
Similarly, in Figure A6, I show the item information curves for a QUIP index that relies on a graded-response model. The monitoring, feedback, and richness components of the measure would hardly contribute to such an index, which would be dominated by the remaining three components instead.

C.5 Empirical distribution of scores

Figure A7 provides descriptive information on the empirical distribution of QUIP scores. Its top panel, which shows histograms for each of the six dimensions, leads to three main observations.

First, on overage, the individual QUIP scores document classrooms that are marked by strong organization and productivity. That is, learning time is maximized, the curriculum is followed, and instruction is densely focused on mathematics and science. At the same time, raters perceived of the instruction to be mostly clear and the content to be free of errors.

Figure A6: *Item information curves for a QUIP index that uses IRT*



Notes. This figure provides item information functions as estimated from a Graded Response Model.

Second, however, this is contrasted by lower scores along dimensions that tap into student-centered instruction. The "monitoring" dimension reveals how lessons rarely elicit evidence of student understanding. Moreover, in more than half of the classes the best rating on the "feedback" dimension was "newcomer", reflecting a lack of scaffolded feedback, feedback loops, and the provision of encouragement (beyond correct vs false). Further, the "richness" dimension reveals shortcomings in teachers' promotion of multiple solutions to solve problems, rich explanations that focus on deeper understanding, and lessons that connect instruction to students' everyday life experiences.

Third, going back to the previous results on the multidimensionality of instructional quality, it is notable that these positive and negative observations align with the two dimensions I identified through a factor analysis. Taken together, this reveals one dimension in which instruction clearly satisfies curricular and managerial demands, and another, in which students remain at the sidelines, do not receive quality feedback, and do not experience instruction that promotes deep learning.

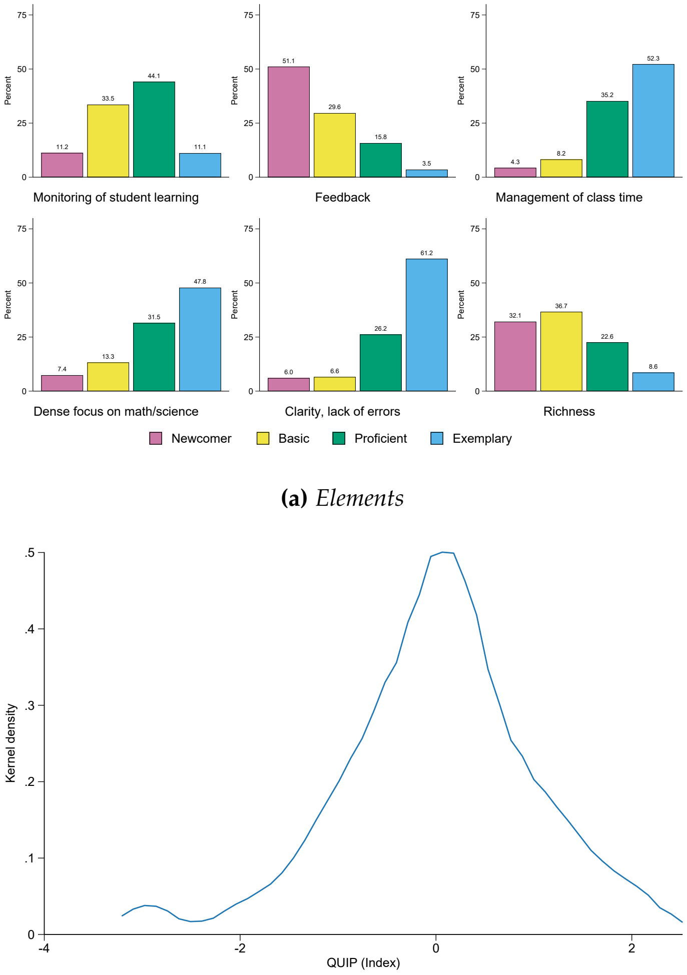
Finally, in the bottom panel, I show a kernel density plot for the index (that is, the inverse covariance-matrix-weighted average, (following Anderson, 2008). The distribution of the index is approximately normal.

C.6 Coherence

C.6.1 Inter-rater agreement

I find satisfactory levels of coherence among the in-person QUIP ratings used in the study. Across the six elements, Cronbach's alpha is 0.77.

Figure A7: *Empirical distribution of QUIP scores*



(a) *Elements*

(b) *Index*

Notes. This figure reports on the distribution of QUIP scores. Subfigure (a) shows histograms for each QUIP element. Subfigure (b) shows a kernel density plot for the index. "Index" refers to the results from an inverse covariance matrix weighted aggregate across the six elements.

I further compare this *overall* level of reliability to inter-rater reliability at the level of invidual "snippets" (recall that observers rate four snippets per lesson, of approximately six minutes each). I make this comparison both across in-classroom ratings (conducted by the NGO), across video ratings (conducted by an external team of raters), and across the in-person and video-based ratings. Table A7 shows the results from this analysis.

Table A7: *QUIP inter-rater reliability*

	In-person			Video			In-person vs video		
	Exact/±0.5	±1	ICC	Exact/±0.5	±1	ICC	Exact/±0.5	±1	ICC
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Elements									
Monitoring of student learning	60.78	93.53	0.58	74.18	100.00	0.68	49.12	88.50	0.32
Feedback	65.09	91.38	0.49	90.16	98.77	0.36	68.58	92.48	0.10
Management of class time	59.45	93.70	0.68	73.78	96.00	0.85	53.40	89.81	0.68
Dense focus on math/science	58.27	90.16	0.64	86.73	97.35	0.90	53.62	90.82	0.63
Clarity, lack of errors	62.90	91.13	0.66	92.34	97.45	0.92	58.90	82.65	0.50
Richness	56.85	89.11	0.44	70.34	97.03	0.75	34.70	78.08	0.09
Panel B: Index									
Mean score	65.98	92.27	0.40	94.05	98.92	0.89	57.22	79.44	0.29

Notes. This table reports on the inter-rater reliability of QUIP scores. "In-person" refers to two in-classroom ratings completed by the NGO's observer and her supervisor. "Video" refers to two video-based ratings completed by external raters. "In-person vs video" refers to one in-person rating completed by the NGO observer and one video-based rating completed by an external rater. Panel A shows results per QUIP element; Panel B shows results for an index. "Index" refers to the mean score across elements. ±0.5 and ±1 refer to the percentage of raters agreeing within 0.5 and 1 points, respectively. For element-wise comparisons, I report on exact matches instead of agreements within 0.5 points. ICC refers to the intraclass correlation coefficient.

The table suggests high levels of agreement if ratings share the same medium of observation. For in-person observations, for the six elements, approximately 90 percent of ratings are within one point (on the four point scale), and about two thirds of observations match exactly. For video-based observations, these numbers are even higher (above 97 percent and above 70 percent, respectively). The agreement of ratings is slightly lower once video-based and in-person ratings are compared to each other. At the snippet-level, the inter-rater reliability as measured by the intra-cluster correlation (ICC) is good for video-based observations and moderate for in-person observations. The ICC points to weaker inter-rater reliability for the "feedback" and "richness" elements, and it is also weaker if video-based and in-person ratings are compared to each other.

C.6.2 Rater effects

In-person classroom observations were administered by the NGO that developed the intervention. Together with the above-mentioned reduction in inter-rater reliability across NGO-based and external ratings, this raises concerns that—beyond differences across in-person and video-based scores—NGO-based ratings may systematically differ in the treatment groups. I investigate, and find support for, this hypothesis in Table A8.

In Table A8, I report on differences in NGO-administered ratings of individual snippets, across the study' experimental groups (overall, and by subject). More specifically, I investigate whether such differences are more (/less) pronounced in the two treatment groups, and report regression coefficients from respective difference-in-difference analyses.

Table A8: *Systematic differences in NGO-administered QUIP ratings, by experimental group*

	Overall		Mathematics		Science	
	ICT	Workbook	ICT	Workbook	ICT	Workbook
	(1)	(2)	(3)	(4)	(5)	(6)
Monitoring of student learning	0.48**	0.38**	0.50*	0.36	0.36	0.35
	(0.20)	(0.18)	(0.26)	(0.23)	(0.25)	(0.26)
Feedback	0.21	-0.02	0.39*	0.19	-0.03	-0.31
	(0.17)	(0.20)	(0.21)	(0.23)	(0.17)	(0.20)
Management of class time	0.71**	0.34	0.73	0.46	0.42	-0.01
	(0.36)	(0.37)	(0.51)	(0.50)	(0.33)	(0.33)
Dense focus on math/science	0.84**	0.06	1.07***	0.45	0.08	-0.84**
	(0.33)	(0.33)	(0.41)	(0.41)	(0.40)	(0.39)
Clarity, lack of errors	0.17	-0.29	0.22	-0.26	0.01	-0.40
	(0.29)	(0.33)	(0.42)	(0.43)	(0.39)	(0.44)
Richness	1.35***	0.31	1.30***	0.39	1.29***	0.15
	(0.24)	(0.23)	(0.30)	(0.28)	(0.35)	(0.34)

Notes. This table investigates whether differences in QUIP scores across in-person ratings (administered by the NGO) and video-based ratings (administered by two external raters) differ by treatment group. Each column compares a treatment group's difference with the difference observed in the control group (i.e., the difference-in-difference). Each cell refers to a separate regression, at the snippet-level. "Overall" pools snippets across subjects; "Mathematics" and "Science" report on results by subject. Coefficients correspond to interaction terms indicating an NGO-administered rating and the treatment groups (ICT or Workbook, respectively). Main differences between in-person and video-based ratings (in the Control group), and main differences across experimental groups (for video-based ratings) are not shown. Standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.

The table suggests that there are systematic differences, with NGO-based ratings largely outperforming those from external ratings, in the treatment groups, especially in ICT schools and for mathematics (the findings for science are mixed).

This finding may be interpreted as systematic bias. It may also be interpreted as evidence that video-based ratings do not capture some aspects of the intervention that improve instructional quality. To allow for either interpretation, my analyses on the effects of the program on instructional quality separately report on the original, NGO-based ratings and on an adjusted version of these ratings. The latter subtracts the coefficients reported in Table A8 (Columns 3 to 6) from each snippet's in-person rating, prior to calculating the standardized QUIP scores and their index.

C.7 Predictive associations

Finally, in Table A9, I investigate correlations between QUIP scores and student test scores, in mathematics and science. I present the results from regressing students' follow-up scores on their class' average QUIP score. I calculate these regressions without controls, after controlling for baseline scores, and after moreover controlling for baseline covariates. In summary, I do not find the expected positive correlations. Instead, for mathematics, I find a negative correlation for maximization of learning time and for whether the presentation of content is clear and free of errors.

Table A9: *QUIP associations with student learning*

	Mathematics			Science		
	Follow-up	Growth	Growth (Controls)	Follow-up	Growth	Growth (Controls)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Mathematics						
Monitoring of student learning	0.12	0.10	0.06			
	(0.08)	(0.08)	(0.07)			
Feedback	0.03	-0.00	0.02			
	(0.07)	(0.07)	(0.06)			
Maximization of learning time	-0.06	-0.04	-0.06			
	(0.07)	(0.07)	(0.05)			
Classroom work is dense	-0.01	-0.01	-0.05			
	(0.06)	(0.06)	(0.06)			
Presentation of content	-0.13	-0.12	-0.13*			
	(0.09)	(0.08)	(0.07)			
Richness	-0.12	-0.12	-0.07			
	(0.09)	(0.08)	(0.08)			
QUIP (Index)	-0.06	-0.07	-0.06			
	(0.07)	(0.07)	(0.07)			
Panel B: Science						
Monitoring of student learning				0.09*	0.04	0.03
				(0.05)	(0.05)	(0.04)
Feedback				0.04	0.03	0.04
				(0.06)	(0.06)	(0.04)
Maximization of learning time				0.01	-0.01	0.01
				(0.06)	(0.05)	(0.05)
Classroom work is dense				-0.02	-0.04	0.01
				(0.07)	(0.05)	(0.05)
Presentation of content				-0.02	-0.05	-0.02
				(0.08)	(0.08)	(0.07)
Richness				0.04	0.00	-0.02
				(0.06)	(0.07)	(0.06)
QUIP (Index)				0.03	-0.00	0.02
				(0.06)	(0.06)	(0.05)

Notes. Each table cell reports the regression coefficient from separate regressions of students' follow-up test scores on QUIP scores, in Control schools. All QUIP scores are aggregated to the mean for a student's school, class, and subject. "Index" refers to the inverse covariance matrix-weighted average, following Anderson (2008). Growth indicates the inclusion of baseline scores as controls. "Controls" indicates the additional inclusion of a vector of student- and school-level covariates, selected via LASSO (see Appendix E). Standard errors in parentheses, clustered at the school level. * significant at 10%; ** significant at 5%; *** significant at 1%.

D Measuring student learning

D.1 Objectives

The study administered tests to measure what students know and can do in mathematics and science, before the intervention was rolled out ("baseline assessment") and thereafter ("follow-up assessment"). Broadly, I followed the "Standards for Educational and Psychological Testing" (American Educational Research Association, 2014); more specifically, I aimed for the assessments to satisfy the following criteria.

First, the tests' content should be narrowly aligned with the official curriculum used in schools, it should measure multiple sub-domains of content knowledge, and it should allow for the measurement of students' knowledge in materials below their enrolled grade-level. Tests should also tap into multiple cognitive domains of varying complexity. Second, the measurement of student ability should allow for students' knowledge to be mapped onto common scales (one per subject), across grades and test occasions. Third, tests should be administered with minimal interference or cheating. Fourth, the tests should measure student ability with high levels of precision, even for students at the extreme tail ends of the ability distribution.

D.2 Test content

I designed the mathematics and science tests to cover a wide range of content domains, grade-level materials, and cognitive complexity. In Section 3.1, I have already given a broad overview of these domains. Here, in Table A10, I provide a more detailed breakdown of the number of questions per domain (see Columns 1 to 10). As shown in the table, I used an approximately even distribution of items across the respective content domains, cognitive domains, and grade-levels.

For both mathematics and science, the test questions drew from a large item pool, from other large-scale assessments. These include, but are not limited to, the Andhra Pradesh Randomized Studies in Education (APRESt), Central Board of Secondary Education (CBSE) board exams, India's National Achievement Surveys (NAS), OECD's Programme for International Student Assessment (PISA), the India-based Student Learning Survey (SLS), and the Trends in Mathematics and Science Study (TIMSS). All items were selected for their alignment with the official CBSE curriculum. They were translated, piloted, and adjusted to the Indian context (if necessary).

There is no reason to believe Avanti Fellows coached students along with the tests ("teaching to the test"). The test content was shared with a separate team at Avanti, which is not responsible for the development or implementation of the intervention. With this team, I confirmed that none of the test questions appear in any of the materials used in the intervention (e.g., in the workbooks distributed to students). Moreover, Avanti Fellows' field staff did not have access to the assessments prior to test administration, and they do not directly teach students in schools.

Table A10: *Number of items per test, skill, and grade-level*

	Grade 9					Grade 10					Anchors	
	Total	Higher-order	Lower-order	At level	Below level	Total	Higher-order	Lower-order	At level	Below level	Across grades	To prev. year
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Mathematics baseline												
Algebra	10			2	8	10			4	6	6	
Geometry	10			3	7	10			4	6	5	
Number sense	10			4	6	10			4	6	6	
Statistics and reasoning	10			5	5	10			4	6	6	
Panel B: Science baseline												
Biology	11			6	5	14			7	7	7	
Chemistry	12			8	4	13			7	6	6	
Physics	12			7	5	13			5	8	8	
Panel C: Mathematics follow-up												
Algebra	8	2	6	4	4	8	2	6	4	4	2	4
Geometry	8	0	8	3	5	8	5	3	4	4	1	4
Number sense	8	3	5	3	5	8	2	6	4	4	2	4
Statistics and reasoning	8	6	2	4	4	8	5	3	4	4	2	4
Panel D: Science follow-up												
Biology	12	2	10	6	6	12	6	6	6	6	2	7
Chemistry	12	6	6	6	6	12	7	5	6	6	3	6
Physics	12	3	9	6	6	12	6	6	6	6	3	6

Notes. This table provides the number of items on the baseline and follow-up assessments. "Anchors" refers to repeat items, across grades (Column (11)) and across test administrations (Column (12)).

D.3 Test booklets

I used multiple test booklets for both subjects—across baseline and follow-up tests, across grades, and within each grade. The follow-up test repeats approximately half of the baseline items, to allow for the linking of test scores across test occasions. During each assessment round, tests are grade-level specific but also share overlapping items, to allow for the linking of test scores across grades. Finally, within each classroom, students were assigned to alternating versions of the test of different question order (sets "A" and "B"), to avoid cheating.

I selected repeated questions ("anchors") according to their item characteristics from pilot and baseline assessments. I moreover aimed for an equal share of anchors across grade-levels, content domains, and cognitive domains. Table A10 summarizes the resulting number of repeated items across grades (Column 11) and test occasions (Column 12).

D.4 Test administration

The baseline and follow-up tests consisted of paper-based assessments that were administered by school-external staff (on December 14, 2018 and November 19, 2019, respectively). The following subsections provide additional information on field operations and quality control.

D.4.1 Field operations

Schools in the study sample were informed two days prior to the assessments. The assessments were then administered by independent government invigilators, at the school level. Government invigilators reported to their assigned schools an hour before the assessment. They carried a school packet which contained, for each grade, Optical Mark Recognition (OMR) sheets, the two sets of question papers, an attendance sheet, and a government-issued authorization letter.[22]

By grade, students were seated in separate examination halls thirty minutes before the assessment started. Government invigilators used a blackboard/whiteboard to demonstrate the method of marking responses using an OMR sheet. Students were given two hours to solve the assessment. Regardless of how soon students finished solving the assessment, they had to remain seated in the examination hall for at least one and a half hours.

After the assessments were completed, government invigilators collected OMR sheets and the question papers. These were packed in the envelopes along with attendance sheets. The principal signed and stamped each envelope. Government invigilators carried these envelopes to the office of District Project Coordinator (DPC). All envelopes were then sent to a central location, for data entry and processing.

[22] Principals also appointed a teacher from the faculty to assist government invigilators and field staff in arranging the logistics of the assessment. To avoid a potential conflict of interest, teachers appointed by the principal taught Hindi, Sanskrit or social science (i.e., not mathematics or science).

D.4.2 Quality control

Quality checks followed procedures similar to those of the state-issued board exams, with additional monitoring. The assessments were administered under the supervision of government invigilators, minimizing the involvement of school faculty. The Assistant Project Coordinator (APC) of every district assigned government invigilators to schools in the study sample. An invigilator-student ratio of approximately 1:50 was maintained.

In addition to government invigilation, a subset of schools was spot-checked in surprise visits (31 schools during the baseline and 84 schools during the follow-up assessment). Spot-checkers consisted of an independent team of field staff, who were expected to visit one school each. For the follow-up test, I calculated an index of potential cheating, using the baseline data and following Jacob and Levitt (2003). Spot-checkers then targeted those schools with the highest expected propensity to cheat, with an equal split across the three experimental groups.

D.5 Scoring

The study's main outcomes are continuous test scores in mathematics and science. I obtain these scores with Item Response Theory (IRT). In secondary analyses, I also investigate whether students are "proficient in" (or "mastered") a given sub-domain on the test. I obtain these classifications of students with Cognitive Diagnostic Models (CDMs). The particular IRT and CDM methods I use in this study rely on students' responses to individual test questions and their grading as either correct or incorrect. In the following sub-sections, I provide additional detail for each of the two analytical approaches.

D.5.1 Continuous scoring using Item Response Theory

There are two challenges for the calculation of student performance levels, and their comparability across grades and test occasions. First, questions differ across grades and test occasions. Second, items also differ in terms of their difficulty and their ability to discriminate student ability. I use Item Response Theory (IRT) to calculate the study's continuous scores of mathematics and science, as it provides a solution to each of these challenges.

IRT exploits the subset of items that appeared on multiple test papers ("anchors") for the linking of estimates onto one common, continuous ability scale.[23] In this study, I calculate scaled scores with a standard, two-parameter logistic (2PL) IRT model, which moreover explicitly models each question's difficulty and its ability to discriminate (Birnbaum, 1968; Samejima, 1973).[24]

Results are then re-scaled to a mean of zero and a standard deviation of one, using the baseline control group as reference (including attritors). I repeat this standardization

[23] I use a concurrent linking approach. See Stocking and Lord (1983) and Kolen and Brennan (2004).

[24] A three-parameter logistic model did not converge.

separately for each grade and subject; in the Control group, students in both grades thus start the study with a mean baseline value of zero, in each of the two subjects.[25]

D.5.2 Determining student proficiency using Cognitive Diagnostic Models

In the study's secondary analyses, I determined student proficiency through Cognitive Diagnostic Models (CDMs). CDMs are multi-dimensional latent-trait models, which were "developed specifically for diagnosing the presence or absence of multiple fine-grained skills or processes required for solving problems on a test" (de la Torre, 2009, 164). This study largely relies on the generalized deterministic inputs, noisy and gate (G-DINA) model for dichotomous items (deălaăTorre, 2011).

As common for CDMs, the G-DINA model requires a theoretically-founded specification of which attributes are expected to contribute to an examinee's probability of answering a given item *j* correctly. This so-called "Q-matrix" lists all items as rows, all attributes as columns, and denotes $q_{ja} = 1$ if attribute *a* is reflected in item *j* (and $q_{ja} = 0$, otherwise). The study's student assessments are explicitly designed to provide this item-to-skill mapping.

In CDMs, the mastery profile of each learner is described by a latent vector of dichotomous entries that each indicate whether an examinee has mastered any attribute; $\boldsymbol{\alpha}^*_{lj} = (\alpha_{l1}, \cdots, \alpha_{lk}, \cdots, \alpha_{lK^*_j})$, where K^*_j denotes the number of attributes captured by item *j*. Conditional on this latent vector $\boldsymbol{\alpha}^*_{lj}$, G-DINA models the probability of an examinee's correct answer for *j*, as a function of item parameters λ_j.

Following deălaăTorre (2011), we may express a respondent's probability of solving an item as

$$P(X_j = 1|\boldsymbol{\alpha}^*_{lj}) = \lambda_{j0} + \sum_{k=1}^{K^*_j} \lambda_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K^*_j} \sum_{k=1}^{K^*_j-1} \lambda_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \lambda_{j12\ldots K^*_j} \prod_{k=1}^{K^*_j} \alpha_{lk} \tag{3}$$

, where λ_{j0} reflects the probability of a correct answer to item *j* for non-masters (the "guessing parameter"), λ_{jk} is the main effect related to having mastered attribute *k*, $\lambda_{jkk'}$ captures the interaction effect for attributes *k* and $k^{'}$, and $\lambda_{12\ldots K^*_j}$ is the interaction effect given mastery of attributes 1 to K^*_j.

Finally, recall that I intend to measure student proficiency on two scales—one that reflects mastery at a student's enrolled grade-level, and one that reflects mastery at a grade-level below. In addition, I investigate students' mastery in multiple content domains. Therefore, I performed the above G-DINA estimations in multiple runs, where each run reflects the estimation of a different grade level, or content domain.[26]

[25] Note that scores cannot be compared across subjects. It is rather meaningless to compare any given score in mathematics with another score in science.

[26] Across these runs, I allowed item parameters to vary.

D.6 Empirical distribution of scores

In Figure A8, I show kernel density plots for the empirical distribution of test scores, for students in the Control group.[27] The distribution of the scores is approximately normal, both for mathematics and science. Importantly, the figure also shows no "bunching" of scores at the tail ends of the distribution—I therefore conclude that the test does not suffer from ceiling or floor effects. In the following sub-section, I further investigate (and find evidence for) the tests' precision, including for students of very low (/very high) ability.

D.7 Item fit and reliability

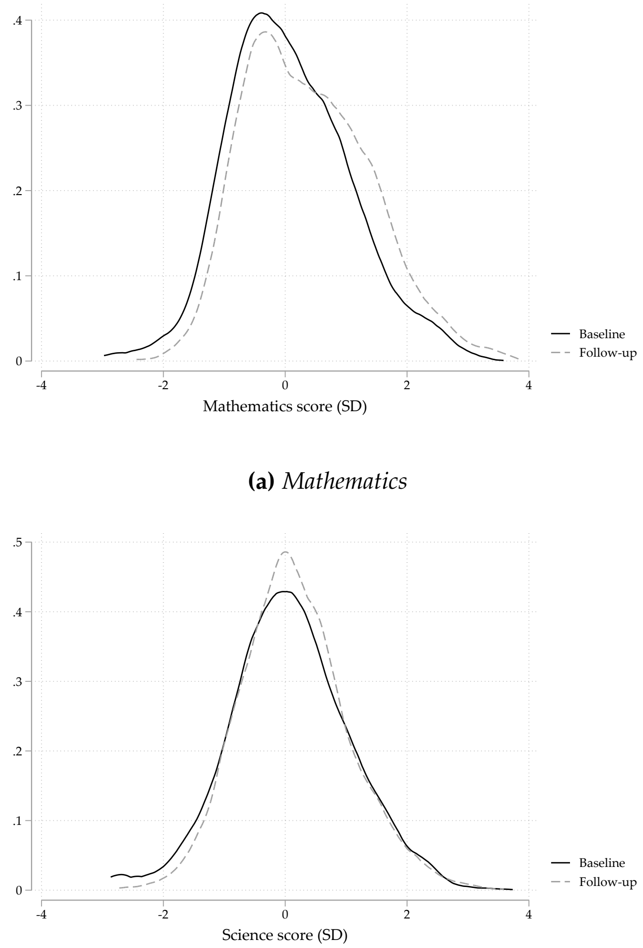
Table A11 and Table A12 provide the discrimination and difficulty parameters for the mathematics and science test questions, as per 2PL IRT models. The table's difficulty parameters show how the tests offer well-distributed measures of student learning in both subjects, as items cover a wide range of difficulty. Moreover, almost all items show high levels of discrimination.

Combined with the test length, these item characteristics translate into high levels of internal consistency. One benefit of item response theory is the ability to report on test precision across a *range* of student ability, not just a single measure of test reliability (such as Cronbach's alpha, for example). This is important as low-ability and high-ability are usually measured with higher levels of noise. Accordingly, I investigate the tests' precision with their test information function (TIF). The information function tells how precisely each ability level is being estimated, along with the corresponding standard error of measurement, at any given level of student ability.

Figure A9 presents the TIF curves for mathematics (top panel) and science (bottom panel), along with the corresponding standard errors. For both subjects, I find high levels of information and low standard errors of measurement, for a wide range of ability. Students two standard deviations below (/above) the median are assessed with a standard error below 0.32 (corresponding to reliability levels above 0.9). Even students three standard deviations below (/above) the median are assessed with a standard error below 0.45 (corresponding to reliability levels above 0.8), even at these extreme levels of student ability.

[27] The figure complements Figure A1, which shows kernel density plots at baseline, across the three experimental groups.

Figure A8: *Empirical distribution of test scores, by subject and assessment round*



(a) *Mathematics*

(b) *Science*

Notes. This figure provides the empirical distribution of test scores, as per 2PL IRT models, for students in the Control group. Each panel shows kernel density plots by assessment round (baseline and follow-up). The top panel reports results for mathematics; the bottom panel reports results for science.

Table A11: *Item characteristics (IRT): Mathematics*

	Mapping		Item parameters (IRT 2PL)		Percent correct			
	Grade-level (1)	Higher/lower (2)	Discrimination (3)	Difficulty (4)	Baseline 9 (5)	Baseline 10 (6)	Follow-up 9 (7)	Follow-up 10 (8)
Algebra								
P021420	9	n/a	0.93	0.24		0.43		
P021500	6	n/a	1.02	0.51	0.41	0.35		
P021504	6	n/a	1.87	-0.33	0.57			
P027952	7	n/a	0.64	-0.53	0.57	0.55		
P027972	8	n/a	0.90	0.53	0.39			
P051137	7	n/a	0.95	0.05	0.47	0.47		
P051143	7	n/a	1.34	-0.69	0.63	0.65		
P060114	9	n/a	0.79	1.03		0.31		
P084845	6	Lower-order	1.53	-0.25	0.51	0.52	0.58	0.65
P084882	7	Lower-order	1.41	0.20	0.43	0.36		0.57
P085370	7	Lower-order	1.47	0.10	0.46		0.47	0.52
P085375	9	Lower-order	1.67	0.01		0.46	0.48	
P085378	9	Lower-order	1.28	0.67		0.32	0.33	
P085521	8	Lower-order	1.08	-0.33	0.45			0.72
P087391	8	Higher-order	1.02	0.48			0.40	
P087595	9	Lower-order	1.01	-0.16			0.54	
P087689	10	Lower-order	1.03	0.47				0.44
P087694	10	Lower-order	0.66	1.96				0.26
P103034	8	Higher-order	0.88	1.01			0.32	
P104677	10	Higher-order	0.34	2.84				0.29
P107067	10	Higher-order	0.70	0.73				0.42
P107070	9	Lower-order	0.87	0.13			0.48	
Geometry								
P021487	6	n/a	1.11	-1.56	0.79			
P021519	6	n/a	0.71	-0.02	0.49			
P027794	7	n/a	1.47	-0.74	0.64	0.67		
P027808	7	n/a	0.33	3.06	0.27	0.26		
P027948	9	n/a	0.93	-0.45		0.56		
P027955	8	n/a	0.73	0.39	0.43			
P027966	9	n/a	0.94	0.13		0.45		
P051138	7	n/a	0.83	-0.67	0.60	0.60		
P084825	7	Higher-order	0.90	-0.43	0.57	0.57		0.58
P084883	8	Lower-order	1.05	0.16	0.46	0.41	0.46	0.55
P085373	9	Lower-order	1.41	-0.50		0.57	0.65	
P085384	9	Lower-order	1.09	-0.08		0.50	0.50	
P085416	8	Lower-order	1.00	0.94		0.30	0.30	
P085502	7	Higher-order	0.83	1.33	0.28			0.29
P085562	8	Higher-order	0.93	-0.27	0.58			0.54
P048385	10	Lower-order	0.26	3.98				0.27
P087395	9	Lower-order	1.25	-0.03			0.51	
P087698	10	Higher-order	0.00	369.53				0.17
P087699	10	Lower-order	0.46	1.29				0.38
P087702	10	Higher-order	0.38	4.35				0.18
P103017	7	Lower-order	0.95	0.07			0.49	
P104343	8	Lower-order	0.97	-0.35			0.57	
P104673	8	Lower-order	0.86	0.64			0.39	
Number sense								
P021206	9	n/a	0.79	-0.19		0.51		
P027816	8	n/a	1.06	-0.25	0.58	0.49		
P027823	7	n/a	1.08	0.30	0.45	0.38		
P027927	7	n/a	1.21	0.03	0.46	0.46		
P027938	8	n/a	0.75	0.25	0.40	0.47		
P027958	7	n/a	0.66	0.30	0.45			
P043873	9	n/a	0.64	0.45		0.42		
P060120	7	n/a	0.50	1.67	0.31			
P084830	8	Higher-order	1.36	-0.35	0.57	0.53	0.64	0.62
P084888	8	Lower-order	0.84	0.90	0.31	0.31	0.35	0.40
P085371	9	Lower-order	0.69	1.28		0.31	0.29	
P085414	9	Lower-order	0.68	0.67		0.36	0.42	
P085515	7	Lower-order	1.32	0.50	0.37			0.41
P085546	7	Lower-order	1.19	0.43	0.39			0.44
P087393	9	Lower-order	0.91	0.23			0.46	
P087396	7	Higher-order	0.95	0.49			0.41	
P087603	6	Higher-order	0.96	0.48			0.41	
P099379	10	Lower-order	0.60	1.28				0.36
P104334	7	Lower-order	1.09	0.52			0.39	
P104674	10	Lower-order	0.54	1.69				0.32
P104675	10	Higher-order	0.75	2.24				0.20
P104684	10	Higher-order	0.58	1.67				0.31
Statistics/Reasoning								
P026937	7	n/a	0.62	-0.15	0.51			
P027804	9	n/a	0.52	1.54		0.31		
P034805	8	n/a	0.84	0.38	0.42			
P038395	8	n/a	0.41	2.44	0.25	0.28		
P043813	9	n/a	1.04	-0.09		0.49		
P051127	7	n/a	0.82	-0.55	0.56	0.59		
P051135	7	n/a	0.98	-1.28	0.72	0.73		
P059470	6	n/a	1.21	-1.73	0.86	0.81		
P084817	8	Lower-order	0.85	-0.25	0.52			0.60
P084836	7	Lower-order	1.28	-0.79	0.66	0.62	0.70	0.77
P084891	8	Higher-order	0.80	0.82	0.38	0.35		0.34
P084893	8	Higher-order	1.25	0.61	0.33		0.37	0.40
P085393	9	Higher-order	0.69	1.92		0.21	0.25	
P085413	9	Higher-order	0.71	1.01		0.33	0.34	
P066086	10	Lower-order	0.63	0.93				0.40
P087405	8	Higher-order	0.77	-0.20			0.54	
P087558	8	Higher-order	0.76	0.49			0.42	
P087717	10	Higher-order	0.35	1.50				0.39
P087719	10	Higher-order	0.87	1.55				0.27
P087722	10	Lower-order	0.49	2.31				0.27
P103021	9	Lower-order	0.75	0.74			0.38	
P107071	9	Higher-order	0.26	5.00			0.22	

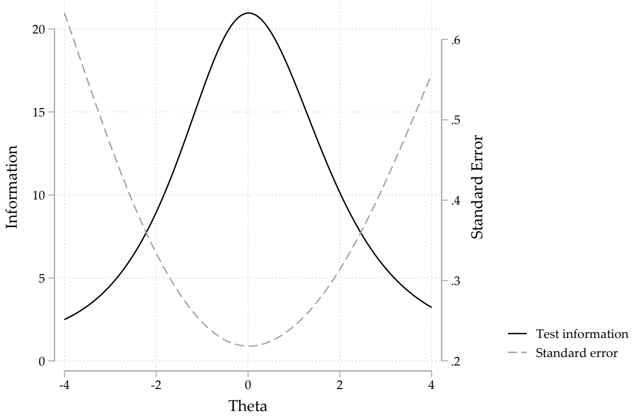
Notes. This table provides item characteristics as per a 2PL item response theory (IRT) model. Items are sorted by content domain. Item names refer to study-internal question IDs. For reference, the table also provides each items' grade-level mapping (Column 1), whether the item is mapped to higher- vs lower-order thinking skills if available (Column 2), and the average percentage of correct answers during the baseline (Columns 5 and 6) and follow-up assessments (Columns 7 and 8).

Table A12: *Item characteristics (IRT): Science*

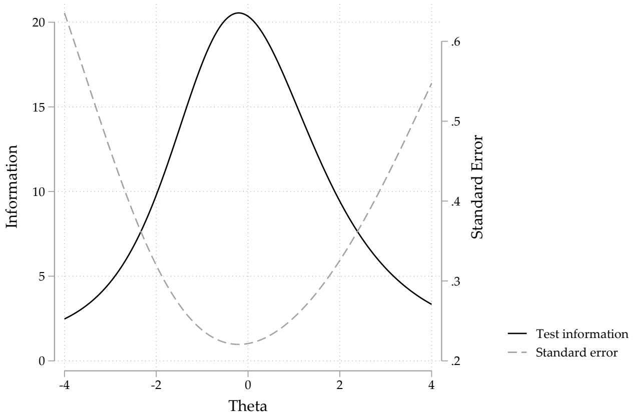
	Mapping		Item parameters (IRT 2PL)		Percent correct			
	Grade-level (1)	Higher/lower (2)	Discrimination (3)	Difficulty (4)	Baseline 9 (5)	Baseline 10 (6)	Follow-up 9 (7)	Follow-up 10 (8)
Biology								
P021356	9	n/a	0.98	0.21		0.43		
P021631	8	n/a	1.06	-0.08	0.50	0.51		
P021634	6	n/a	1.19	0.29	0.44			
P021643	6	n/a	1.05	-0.04	0.51			
P022088	9	n/a	0.70	0.34		0.43		
P026003	9	n/a	0.71	0.66		0.38		
P026025	7	n/a	0.79	-0.01	0.51	0.48		
P054655	9	n/a	0.76	1.17		0.30		
P084959	7	Lower-order	0.91	1.29	0.28			0.27
P084961	8	Lower-order	1.06	-0.17	0.52	0.50	0.53	0.59
P084963	8	Lower-order	1.55	-0.56	0.64	0.60	0.65	0.71
P084973	8	Lower-order	0.75	0.86	0.35	0.34	0.38	
P085428	9	Lower-order	0.93	-0.23		0.52	0.56	
P085429	9	Lower-order	1.35	-0.91		0.68	0.74	
P085431	9	Lower-order	0.63	1.58		0.27	0.28	
P085536	8	Higher-order	0.88	0.55	0.45	0.36		0.39
P085537	8	Higher-order	0.70	1.48	0.27	0.28		0.29
P085538	7	Lower-order	0.82	0.40	0.45			0.43
P051367	10	Lower-order	1.17	0.49				0.41
P086914	8	Lower-order	1.09	-0.13			0.53	
P087469	9	Lower-order	0.79	0.05			0.49	
P087471	7	Lower-order	1.21	-0.06			0.52	
P087475	9	Higher-order	0.12	10.80			0.22	
P087659	10	Higher-order	0.87	1.15				0.31
P087661	10	Lower-order	1.32	-0.47				0.65
P087663	10	Higher-order	1.05	0.72				0.36
P087666	10	Higher-order	1.29	0.68				0.35
P087667	10	Higher-order	1.27	0.93				0.30
P103037	9	Lower-order	0.37	1.76			0.35	
P103046	7	Higher-order	1.07	-0.36			0.58	
Chemistry								
P021392	9	n/a	0.69	2.00		0.21		
P021399	9	n/a	0.88	0.66		0.36		
P021401	9	n/a	0.87	0.88		0.32		
P021628	6	n/a	0.71	0.37	0.45			
P025446	8	n/a	0.92	0.27	0.48	0.40		
P025450	8	n/a	1.29	-0.44	0.60	0.58		
P025452	8	n/a	0.51	0.35	0.48	0.43		
P039159	6	n/a	1.42	-0.13	0.54			
P056305	7	n/a	0.99	0.08	0.49			
P072312	9	n/a	0.56	2.43		0.21		
P084976	8	Higher-order	1.12	0.38	0.42	0.36	0.40	0.50
P084986	8	Lower-order	1.35	-1.17	0.77	0.74	0.76	0.81
P084991	8	Higher-order	1.40	-0.52	0.62	0.58	0.64	0.70
P085044	7	Higher-order	0.94	-0.04	0.53			0.50
P085433	9	Lower-order	1.39	-1.05		0.72	0.76	
P085435	9	Lower-order	1.12	-0.07		0.51	0.49	
P085437	9	Higher-order	1.11	-0.03		0.49	0.49	
P085539	8	Lower-order	0.98	0.20	0.43			0.51
P085540	8	Higher-order	0.62	1.44	0.33			0.29
P086928	8	Higher-order	1.29	0.46			0.39	
P086932	7	Higher-order	0.89	0.76			0.36	
P087409	9	Lower-order	0.63	1.21			0.33	
P087410	9	Higher-order	1.49	-0.63			0.67	
P087412	8	Lower-order	1.53	-0.38			0.60	
P087676	10	Higher-order	0.64	1.15				0.35
P087677	10	Higher-order	1.20	0.70				0.36
P087678	10	Lower-order	0.95	0.43				0.43
P087680	10	Lower-order	0.72	1.17				0.33
P087681	10	Higher-order	0.90	0.81				0.36
P087683	10	Lower-order	0.74	1.22				0.32
P103043	9	Lower-order	0.32	0.98			0.42	
Physics		n/a						
P013841	7	n/a	0.73	-0.01	0.53	0.46		
P021822	9	n/a	0.53	1.40		0.32		
P024152	8	n/a	0.65	-0.16	0.53	0.50		
P024908	8	n/a	0.53	1.67	0.30	0.30		
P025624	8	n/a	0.56	0.78	0.40			
P054668	7	n/a	0.75	0.07	0.44	0.51		
P054831	7	n/a	0.52	1.18	0.35	0.36		
P059488	9	n/a	0.73	1.14		0.31		
P084898	7	Higher-order	1.28	-1.09	0.74	0.71	0.78	0.77
P084900	8	Higher-order	0.77	0.06	0.50	0.46	0.50	0.51
P084906	8	Lower-order	0.74	-0.31	0.59	0.53	0.56	0.52
P084953	8	Lower-order	1.23	-0.52	0.62			0.66
P085419	9	Lower-order	0.99	0.17		0.43	0.48	
P085426	9	Higher-order	0.88	0.06		0.48	0.48	
P085427	9	Lower-order	0.56	1.00		0.36	0.38	
P085534	7	Lower-order	1.03	-0.01	0.51			0.52
P085535	8	Higher-order	0.52	1.65	0.35			0.28
P087419	8	Lower-order	1.22	-0.43			0.60	
P087420	8	Lower-order	0.44	0.59			0.44	
P087668	10	Lower-order	0.66	1.73				0.27
P087670	10	Lower-order	0.77	0.15				0.49
P087679	10	Lower-order	1.10	1.17				0.27
P087684	10	Higher-order	0.90	0.74				0.38
P087685	10	Higher-order	0.55	1.94				0.28
P087686	10	Higher-order	0.36	5.21				0.14
P103018	8	Lower-order	0.90	0.69			0.37	
P103036	9	Lower-order	0.45	2.56			0.25	
P103049	9	Lower-order	0.88	-0.70			0.63	
P104335	9	Lower-order	1.00	0.48			0.40	

Notes. This table provides item characteristics as per a 2PL item response theory (IRT) model. Items are sorted by content domain. Item names refer to study-internal question IDs. For reference, the table also provides each items' grade-level mapping (Column 1), whether the item is mapped to higher- vs lower-order thinking skills if available (Column 2), and the average percentage of correct answers during the baseline (Columns 5 and 6) and follow-up assessments (Columns 7 and 8), by grade.

Figure A9: *Test information functions (TIF)*



(a) *Mathematics*



(b) *Science*

Notes. This figure provides the test information functions, and corresponding standard errors of measurement, for the mathematics (top panel) and science (bottom panel) tests, as per 2PL IRT models.

E Identifying a vector of controls

As potential covariates, I considered all of the paper's baseline data sources, including village-/town characteristics, school characteristics, student characteristics, and results from the baseline assessment (see Section 3.1). From these data, I excluded those variables without information for all students or schools.

To select a vector of control variables, I then implemented the post double Least Absolute Shrinkage and Selection Operator (LASSO), following Belloni *et al.* (2014). For simplicity, I identified a common set of controls, for all estimations. To do so, I focused on a simple average across students' follow-up mathematics scores (2PL, std.) and science scores (2PL, std.). More specifically, the LASSO procedure uses residuals from a regression of this average on treatment group indicators and randomization strata fixed effects.

Table A13 below lists the set of potential variables and whether they were selected as covariates, or not. In models where the outcome variable is not at the student level (e.g., outcomes measured through classroom observations), I control for the school-level average of the selected variables.

Table A13: *Covariate selection using LASSO*

	Selected (1)
Baseline assessment	
Grade	Yes
Mathematics score (2PL, std.)	Yes
Science score (2PL, std.)	Yes
Math score squared (2PL, std.)	No
Science score squared (2PL, std.)	No
Mathematics percent correct	No
Science percent correct	No
Mastery of algebra, at grade-level	No
Mastery of geometry, at grade-level	Yes
Mastery of number sense, at grade-level	No
Mastery of statistics/reasoning, at grade-level	Yes
Mastery of biology, at grade-level	No
Mastery of chemistry, at grade-level	No
Mastery of physics, at grade-level	No
Mathematics, below level	No
Mathematics, at level	No
Science, below level	No
Science, at level	No
Mastery of alg, below grade-level	Yes
Mastery of alg, at grade-level	No
Mastery of geo, below grade-level	No
Mastery of geo, at grade-level	Yes
Mastery of num, below grade-level	No
Mastery of num, at grade-level	No
Mastery of str, below grade-level	Yes
Mastery of str, at grade-level	Yes
Mastery of bio, below grade-level	No
Mastery of bio, at grade-level	No
Mastery of chm, below grade-level	Yes
Mastery of chm, at grade-level	Yes
Mastery of phy, below grade-level	No
Mastery of phy, at grade-level	Yes
Mastery of mathematics, below grade-level	No
Mastery of mathematics, at grade-level	No
Mastery of science, below grade-level	Yes
Mastery of science, at grade-level	Yes
DISE	
School: years in service	No
School is co-ed (vs. single-sex)	No
Percentage of classrooms needing minor repair	No
Percentage of classrooms needing major repair	No
No. toilets / students	Yes
Boundary wall is inexistent or incomplete	Yes
School has tap water	No
Computers / no. of students	No
Received a school development grant	No
Received a school maintenance grant	No
Infrastructure audit: Available	
Projector	No
Remote	No
Screen	No
Speakers	No
Computer	No
Internet	No
Generator	No
Cupboard	Yes
Infrastructure audit: Functional	
One functioning smart classrooms or more	Yes
Two functioning smart classrooms or more	No
Projector	No
Remote	Yes
Screen	Yes
Speakers	No
Computer	No
Internet	No
Generator	No
Cupboard	No
Board exams	
Total number of students	No
Average score, mathematics	No
Average score, science	No
Percentage failing, mathematics and science	No
Percentage failing, overall	No
Percentage above 50, overall	No
Percentage above 60, overall	Yes

Notes. This table reports on the selection of control variables, from baseline student and school characteristics. The outcome variable is the residuals from a regression of the average across students' follow-up mathematics score (2PL, std.) and science score (2PL, std.) on treatment group indicators and randomization strata fixed effects. Potential covariates without information for all students / schools were excluded (not shown here). Selection uses LASSO, following Belloni *et al.* (2014). "Yes" indicates that a variable was selected.