Essays on Teacher Effectiveness and Student Learning in Less-Developed Countries

Dissertation Advisor: *Author:*

Felipe Barrera-Osorio **Andreas de Barros**

Essays on Teacher Effectiveness and Student Learning in Less-Developed Countries

Abstract

It has now been widely documented that learning levels in many less-developed countries remain low. While this Global Learning Crisis is increasingly recognized, we still know very little about its specific scale. Moreover, although teacher and teaching quality are considered important determinants of student learning, there is a dearth of knowledge as to how to improve instructional quality through teacher-level interventions.

This dissertation seeks to shed light on these issues through three independent essays. The first essay presents detailed, representative, and previously unpublished, learning outcomes data from India. The second essay investigates the causal effects of repeat, formative performance evaluations, under Chile's national teacher evaluation system. In the third essay, I conduct a cluster-randomized trial in India to investigate the effects of a computer-assisted educational program that encourages teachers to blend their instruction with high-quality video materials.

Contents

List of Figures

List of Tables

Acknowledgments

I sincerely thank my professors for their outstanding support during my doctoral studies. I am particularly grateful to Felipe Barrera-Osorio, my dissertation chair and mentor. Many other professors have supported me throughout: My advisors Karthik Muralidharan and Martin West, as well as Emmerich Davies, Alejandro Ganimian, Heather Hill, Andrew Ho, Paola Uccelli, and Sarah Dryden-Peterson. I could not have wished for better teachers. Many thanks also to my fellow PhD students and friends, in particular Reid Higginson.

For Chapter 1, I acknowledge the contributions made by my co-author Alejandro Ganimian. We thank Karthik Muralidharan for his contributions to this research in its early stages. We gratefully acknowledge the funding provided by the Douglas B. Marshall, Jr. Family Foundation for this research. We thank Andrew Ho, Jalnidh Kaur, Pranav Kothari, Aarthi Muralidharan, Sridhar Rajagopalan, Maulik Shah, Nishchal Shukla, Gayatri Vaidya, and attendants of AEFP's 43rd Annual Conference, for useful feedback that informed this study. We thank Educational Initiatives for making data available. Google.org supported Educational Initiatives with funding for the Student Learning Study (SLS), which provides our source of data. We also thank Anuja Venkatachalam for providing excellent research assistance.

For Chapter 2, for their helpful comments, I would like to thank Clément de Chaisemartin, Olivia Chi, José Ignacio Cuesta, Melissa Dell, David Deming, Pierre de Galbert, Kathryn Gonzalez, Heather Hill, Francisco Lagos, Anne Lamb, Cristián Larroulet, María Lombardi, Eduardo Montero, Abhijeet Singh, and Ugo Troiano. I thank the Chilean Agencia de Calidad de la Educación and the Ministry of Education for making data available.

For Chapter 3, I gratefully acknowledge the Government of Haryana and Avanti Fellows for their contributions to the "Sankalp" project, and for their commitment to evidence-based education policy. I thank Panchali Dutta and Akshay Saxena in particular, and many other Avanti staff members (including Bijeta Mohanti, Vaibhav Shukla, and Pritesh Shrivastava). I am very thankful for outstanding 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 Alejandro Ganimian for his ongoing 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 providing my research with funding. The Michael and Susan Dell Foundation supports Avanti Fellows with funding for the "Sankalp" project and its evaluation.

The usual disclaimer applies, and I have no conflicting interests to declare.

To my Tiemanns.
Danke.
To Divya and the de Barroses.
Merci.

Introduction

Despite great improvements in student enrollment and attendance, it has now been widely documented that learning levels in many developing countries remain low. While this Global Learning Crisis is increasingly recognized, we still know very little about its specific scale. Moreover, although teacher and teaching quality are considered important determinants of student learning, there is a dearth of knowledge as to how to improve instructional quality through teacher-level interventions.

This dissertation seeks to shed light on these issues through three independent essays.[1] The first essay (Chapter 1) presents detailed, representative, and previously unpublished, learning outcomes data on 101,084 public-school students across 18 Indian states and one union territory. I find that primary- and middle-school students perform worse on foundational skills not captured by commonly used assessments, that students show less growth on these skills across grade levels, and that girls perform below boys for these previously unassessed skills.

The second essay (Chapter 2) investigates the causal effects of repeat, formative performance evaluations, under Chile's national teacher evaluation system. The study's main results suggest that student learning, teacher beliefs and teaching behaviors are not positively affected when evaluations are mandated, both in the year of the evaluation and in the year thereafter. These findings rest on data-sources with unusually comprehensive coverage of a national education system—positive effects on student performance can thus be ruled

[1] This is a three-paper dissertation. Each of its three chapters reflect stand-alone research articles. To maintain each article's independence, I may repeat information across articles.

out precisely. The results are not driven by a teacher's level of work experience, by student sorting, by systematic attrition, or by the article's model specification.

In the third essay (Chapter 3), I conduct a cluster-randomized trial to investigate the effects of an intervention that provides teachers with continuous training and materials, encouraging them to blend their instruction with high-quality video materials. In the state of Haryana, India, I randomly assigned 80 out of 240 public schools to this program. I randomly assigned another 80 schools to a "low-tech" version of the program, which removes any program components related to educational technology. In the short run, I document negative program effects on student learning in grades 9 and 10 in mathematics, and no effects in science. I also find detrimental effects on observed instructional quality, and on student perceptions and attitudes towards mathematics and science. These impacts appear to be at least partially driven by the program components that promote blended instruction.

My findings suggest that the extent of the learning crisis in the Indian education system may have been severely underestimated. At the same time, at least in the short run, two promising interventions that aimed to improve instructional quality did not lead to improvements in student learning. Taken together, this dissertation thus serves as a wake-up call for education researchers and practitioners, highlighting both the urgency to improve teacher effectiveness in less-developed countries, and the lack of knowledge to do so.

Chapter 1

Assessing the Global Learning Crisis: A Fine-Grained Diagnosis Using Large-Scale Data from India[1]

1.1 Introduction

It has now been widely documented that learning levels in developing countries are low (UNESCO, 2015; The World Bank, 2017b). Yet, while this "Global Learning Crisis" is increasingly recognized, we still know very little about the specific skills that confront children with greatest difficulty. Much of the "diagnosis" rests on either aggregate test scores, or on short assessments that capture a limited scope of cognitive domains, only. In this article, we investigate whether more nuanced measures of student ability are worth pursuing instead, to assess the Global Learning Crisis.[2]

[1] Academic article to be co-authored with Alejandro Ganimian.

[2] Another question asks whether assessments can provide more detailed information, to inform remedies. The incomplete understanding of children's learning levels may have limited our ability to effectively improve pedagogy, either by targeting student misconceptions (Confrey, 1990; Metcalfe, 2017; VanLehn, 1990), through formative assessment (Kingston and Nash, 2011; Briggs *et al.*, 2012; McMillan *et al.*, 2013), or through differentiated, personalized instruction (Muralidharan *et al.*, 2019). While we do not attempt to evaluate these interventions and their effectiveness in this paper, we recognize their demand for a more detailed understanding of student learning—therefore, our paper also provides a "proof of concept", as to whether such information

Would the diagnosis of a Global Learning Crisis differ, were more nuanced assessments used instead? We recognize scenarios in which the use of nuanced, diagnostic assessments should be rejected, for not satisfying the parsimony principle: A simpler instrument should be used if it leads to similar substantive conclusions. In contrast, however, more nuanced measures should be preferred if they revealed a different extent of the crisis (i.e., previous assessments of the crisis were incomplete or inaccurate), or if they unmasked heterogeneity in prevalence (i.e., the crisis does not affect certain subgroups). In particular, it would be worthwhile understanding if, for individual skills, (1) the observed lack of student learning was less (or more) predominant (especially among those skills not captured by commonly used assessments), (2) if learning deficiencies decreased (or increased) among students in higher grades, and (3) if certain subgroups (geographic, or by gender) were affected differently, by the crisis. We address each of these three points in this article.

In this study, we estimate fine-grained skill profiles for students in Indian public schools, at the individual level. More precisely, we develop a Cognitive Diagnosis Model (CDM) to estimate the extent to which students in grades four, six, and eight have mastered five mathematical skills: “Fractions and Decimals”, “Measurement”, “Number Concepts and Number Theory”, “Operations on Whole Numbers”, and “Shapes and Geometry”. Importantly, our psychometric approach allows us to map students’ ability levels onto a common scale, and we can thus determine whether any of the students (including those enrolled in higher grades) have mastered the same fourth-grade level understanding. In addition, we take advantage of demographic information on examinees, to address the question of whether manifestations of student learning gaps are uniform across geography and gender.

Our analyses rest on representative data from a country which has featured prominently in previous assessments of the learning crisis. These data are hitherto unpublished in academic outlets and they are of unusually extensive coverage. Our study covers 18 of India’s largest states and one union territory (as per the 2011 Indian census, this geographic area

can be gleaned from common assessments.

counted with approximately 861.2 million inhabitants, for about 12.3 percent of the world's population). Another key characteristic of these assessment data is their explicit mapping of exam questions to individual skills (and the availability of item-level information)—often, researchers only have access to student-level scores. Finally, two of the skills (fourth-grade "Number Concepts and Number Theory" and fourth-grade "Operations on Whole Numbers") relate to those skills commonly discussed in assessments of the learning crisis—the remaining three skills are usually either ignored, or not reported on individually.

The study's main findings suggest that measures of number sense and arithmetic alone should not be considered acceptable proxies of mathematical skill—especially not when student ability is compared across grades. We find that eighth-graders' number sense and knowledge of whole number operations is superior to the level of skill found among their fourth-grade and sixth-grade peers, respectively. Mastery of fourth-grade material increases from 43 percent in grade four, to 50 percent in grade six and 61 percent in grade eight. However, this pattern is not observed for the remaining three skill areas. The share of fourth-graders who have mastered fractions, measurement, and geometry, is approximately half the proficiency rate observed for number sense and arithmetic (22 percent vs. 43 percent). We further estimate that only 28 percent of sixth-graders and only 29 percent of eighth-graders have mastered a fourth-grade level understanding of fractions, measurement, and geometry.

Our results moreover point to substantial differences across India's states. For instance, we find that the percentage of masters among Kerala's *sixth*-graders exceeds that of *eighth*-graders in all other states but one; whereas, in contrast, *eighth*-graders in Jammu and Kashmir show a lower share of masters in comparison to all of the remaining states' *fourth*-graders. Across low- and high-performing states, these differences in average mastery are substantial; from a range of 47 percentage points (for fractions and decimals, among eighth graders) up to a range of 75 percentage points (for number sense, among fourth-graders).

Lastly, our estimates suggest that, on aggregate, male students outperform their female peers, for all five skills. Yet, the article's results point to the importance of dis-aggregation, with gender gaps being especially prevalent in Jharkhand, for example, and less pronounced in Karnataka and Punjab.

Taken together, this study thus points to low levels of student learning overall (in line with previous research and the notion of a "learning crisis"). Existing measures severely underestimate the extent of the crisis, however, by focusing on select skills, only. We also find that levels of mastery vary strongly among geographic units (i.e., states), and in terms of learning gaps between male and female students.

The remainder of the paper proceeds as follows. The next section reviews previous academic work on learning levels and profiles in the developing world. The subsequent section provides a short description of the Indian context. This section is followed by a discussion of the study's sample and an introduction to CDMs, including more detailed information on modeling choices and our psychometric analyses. This section also includes the paper's results, presenting skill profiles of students by grade-level, geographic location, and gender. The last section concludes.

1.2 Prior research

This section of the paper provides a short review of existing evidence on learning levels and profiles in the developing world.[3] In summary, we observe a sharp increase in measures of student learning, which document low learning levels and flat learning trajectories. However, we also note that there is a lack of evidence with respect to those individual skills students lack the most.

In recognizing that progress in student enrollment has not coincided with learning gains, policy makers have increasingly shifted their focus to outcome measures and large-scale assessments (Birdsall *et al.*, 2016; Brookings, 2016; Hanushek and Woessmann, 2015). Developing countries now participate in a growing number of international assessments,

[3] Note that some authors (e.g. Kaffenberger and Pritchett, 2017; Rolleston, 2014; The World Bank, 2017b) refer to a student's progress over time with the term "learning profile". With "learning profile" we refer to a student's level of mastery for a given set of skills. We prefer the term "learning trajectory" to describe a student's progress over time.

going beyond measures of "mere" youth and adult literacy. These include international comparisons (e.g., PIRLS and TIMSS), assessments driven by international organizations, such as the OECD (D-PISA) and the World Bank (GPE, READ, SABER), regional efforts (e.g., LLECE, SACMEQ, PASEC, SEA-PLM), and initiatives covering a smaller number of countries, such as ASER (in India and Pakistan), UWEZO (in Kenya, Tanzania, and Uganda), and Young Lives (in Ethiopia, India, Peru, and Vietnam).[4] Moreover, the past decades have also seen a spike in national assessments; while 8 developing countries conducted at least one national assessment in 1990, this number had grown to 64 by 2013 (UNESCO, 2015, 190 et sq.), and in 2015, the National Learning Assessment Mapping Project identified 307 national assessments in 85 mostly low- and lower-middle income countries (EPDC, 2015).

These assessment data have led to at least three observations. First, while highly varied, overall learning levels in developing countries are low, *in relative terms*.[5] For example, using data for 13 African countries, Sandefur (2018) finds that students perform below the 5th percentile in most developed countries.[6] Second, learning gaps are also observed *at the absolute level*, i.e. with respect to official curricula and learning goals. For India's rural districts, for instance, ASER (2017) reports that only 42.5 percent of grade three students can read a grade one text, and only 27.2 percent can do a 2-digit subtraction (in 2016).[7] Third, at least with the beginning of second grade, most *student learning trajectories* remain flat. For example, Filmer *et al.* (2006, 5) argue that these trajectories are "not steep enough" and that

[4] See Birdsall *et al.* (2016), for an overview. The use of common test items across countries (e.g., EGMA and EGRA, or from the aforementioned assessments) has also spurred growth in international comparisons that link samples through item-response theory (Das and Zajonc, 2010; Sandefur, 2018; Singh, 2019).

[5] However, some countries have made exceptional progress over time, such as Vietnam. See Dang and Glewwe (2018) and Singh (2019). See Kaffenberger and Pritchett (2017), for an analysis of how, in ten developing countries, literacy learning is varied but low. For an overview of student learning levels in South Asia, see Dundar *et al.* (2014).

[6] Similarly, focusing on 12-year-olds, Singh (2014), finds that only about half of Ethiopian children and about a quarter of children in India and Peru reach TIMSS' "low achievement" benchmark for grade four (i.e., for 10-year-olds)—in contrast, this number is seven percent in the UK and the US, and three percent in Singapore and Hong Kong.

[7] Respective statistics are also reported for a diverse range of other countries, including those who have achieved (near-)universal enrollment (Filmer *et al.*, 2006; The World Bank, 2017b).

the Millennium Development Goals should be replaced with learning goals instead.[8]

Even though these overall learning gaps are now widely recognized, there is surprisingly little research with respect to the more fine-grained domains on which students lag behind most. A few preliminary attempts should be mentioned here. A first group of research presents evidence for individual test items (referring to them as examples of basic competence) and provides the percentage of students who solve a particular item correctly.[9] For instance, as mentioned above, ASER tests in Pakistan and India (ASER, 2017) report on the percentage of students who can correctly subtract two-digit numbers, with borrowing, and the percentage of students who can read a grade-one text.[10] Such analyses may also compare the item-wise performance to benchmark across countries.[11] Slightly more complex analyses take a similar approach, yet moreover relate the overall test score to whether a given item is answered correctly (e.g. Andrabi *et al.*, 2008).

A second body of literature reports results by subskill; these reports group questions within topic groups and thereafter report the percent correct or overall test score for this item-group. For example, Sri Lanka's National Assessment of Achievement in Grade 4 Pupils reports percent-correct per subskill (NEREC, 2009), and report cards for India's National Achievement Survey include a total score per skill area (MHRD, 2016). Further, (rather arbitrary) cut-off rules have been used to classify groups of items as either problematic or unproblematic (e.g. Wonu and Zalmon, 2017). As with analyses that focus on individual items, items may also be grouped as they are related to an overall test score (i.e., according to

[8] More recently, Kaffenberger and Pritchett (2017) reiterate this point, and there is now ample empirical, longitudinal evidence to support this observation, with India being one of the most problematic countries (Muralidharan and Zieleniak, 2013; Muralidharan *et al.*, 2019; Pritchett and Beatty, 2015; Rolleston, 2014).

[9] For writing and reading, these tasks may be broader and may refer to the ability to read a simple text.

[10] As another example, Cambodia's national assessment of six graders includes a report on the percentage of students that can write a letter to apologize to someone (MoEYS, 2015).

[11] USAID (2017) presents comparative results from 17 countries, on whether second grade students can answer at least one listening comprehension question or read at least one word of connected text, for example. See Dundar *et al.* (2014, 96), for an example from India, comparing two Indian states with the overall sample of TIMSS countries.

their difficulty level), leading to so-called item maps.[12] Along with an item group's position on these "maps", readers may then use a student's overall, unidimensional score to gauge whether a certain skill should have been mastered or not (cf. Sadler, 1998). With either of these two approaches, however, items are expected to relate to single subskills. Moreover, with these methods, students who attend different grade-levels cannot be compared on a common ability scale.

There are only few examples of research in developing countries that provide direct estimates of examinees' latent skill profiles. These exceptions include early research by Tatsuoka *et al.* (2004), who compared patterns of 23 specific content knowledge and processing subskills, in mathematics, for 20 countries who participated in the 1999 round of TIMSS. This analysis mainly focuses on developed countries, but also includes Chile, Indonesia, and the Philippines. Tatsuoka *et al.* (2004) finds that these countries perform at the bottom for all presented subskills but does not offer a country-by-country discussion of sub-skills. Lee and Sawaki (2009), moreover apply three types of diagnostic models to response data from the internet-based version of the Test of English as a Foreign Language (ToEFL). The study (ibid.) includes examinees in China, Colombia, and India, though results are not broken down by country. The authors find that few learning profiles are mixed (i.e., they are either fully developed or undeveloped), but the study's main focus is on the technical assessment of modeling approaches. Finally, China stands out as the only non-high income country where cognitive diagnostic tests have been implemented at scale, within applications of computer adaptive testing (Liu *et al.*, 2013, 2014). Yet, we are not aware of any related publications that would highlight student skill profiles for low-income students, in China.

1.3 Setting

India's context provides a fruitful setting for a diagnostic (re-)assessment of the Learning Crisis. The country's student population has featured prominently in descriptions of the

[12] See MoPME (2014) for example item maps, from Bangladesh.

crisis, due to its size, its near-universal enrolment levels, and because of low learning levels. As of 2017, India has the world's largest population of children of compulsory school age (UNESCO Institute for Statistics, 2018) and its population of children age 14 or under (372 million) far outstrips that of other countries.[13] By 2002, the year in which India passed a constitutional amendment to recognize education as a fundamental right, the country's net enrollment[14] stood at 79 percent (The World Bank, 2018b). Similar to other developing countries, over a relatively short period of five years, India further improved upon this number, reaching 91 percent in 2007 and maintaining similarly high levels since.[15] In 2009, India's Right of Children to Free and Compulsory Education Act made education from grades one to eight freely accessible for all children between the ages of six to fourteen. However, previous literature has widely documented the country's low levels of student learning, whether in relative terms, in absolute terms, or with respect to students' learning trajectories over time (see Section 1.2, above).

India's student assessment landscape also reflects the globally heightened interest in outcome measures and large-scale assessments. Internationally, India participated in the 2009 round of the Programme for International Student Assessment (PISA, with two states), and will do so again in 2021. Nationally, since 2001, more than three million students participate in India's sample-based "National Achievement Survey" (NAS), every year (covering grades three, five and eight, in science, social science, mathematics and language). At the state level, additional "State Learning Achievement Surveys" (SLAS) mirror the NAS. Neither of these two assessments report student learning profiles. Individual states have moreover introduced "Continuous and Comprehensive Evaluations" (CCE), testing students as often as monthly, on short curricular units.[16] Finally, with ASER, one of the

[13] This includes China (ranked two, at 245 million), Nigeria (ranked three, at 84 million), or the United States (ranked six at 62 million) (The World Bank, 2018b).

[14] "Net enrollment" refers to the ratio of children of official school age who are enrolled in school to the population of the corresponding official school age.

[15] For comparison, this level reflects the OECD members' average net enrollment, in the 1970s.

[16] A recent large-scale, randomized evaluation for the state of Haryana describes this effort as a "failure", ruling out even small positive effects (Berry *et al.*, 2020).

most frequently used instruments to measure the extent of the learning crisis was developed in India. As a potential shortcoming, the instrument focuses on two sub-skills, only (fourth-grade “Number Concepts and Number Theory” and fourth-grade “Operations on Whole Numbers”).

Finally, some of the more promising approaches to tackle the learning crisis also originated in India. Over the past two decades, there have been multiple attempts to address the country’s wide heterogeneity in student preparation (accompanied with randomized evaluations thereof). One strategy found to be effective across different contexts provides targeted remedial education to low-performing children (Banerjee *et al.*, 2007, 2010).[17] Another approach that has been consistently found to be effective in India entails assessing children’s basic skills, regrouping them according to their performance, and providing differentiated activities to each group, periodically retesting children to allow them to “graduate” from one level to the next (“Teaching at the Right Level (TaRL)”, see Banerjee *et al.*, 2017a).[18] As a third approach, in India, computer-assisted learning (CAL) has shown potential to complement regular instruction and successfully provide personalized learning to students (Banerjee *et al.*, 2007; Muralidharan *et al.*, 2019).[19] In spite of their differences, each of these three types of attempts to either accommodate or counteract the heterogeneity of student ability requires a—so far rather crudely implemented—diagnosis of student skill profiles.[20]

[17] See Saavedra *et al.* (2017), for an evaluation of a similar program, in Peru.

[18] TaRL is also consistent with evidence from developing countries with similar characteristics (Duflo and Kiessel, 2014; Duflo *et al.*, 2011).

[19] Most CAL software products that have been rigorously evaluated in other developing countries have only produced small to moderate improvements in student learning (see, for example, Linden, 2008; Carrillo *et al.*, 2010a; Lai *et al.*, 2016, 2013; Mo *et al.*, 2013). One possible explanation for this pattern of results is that these products have largely failed to leverage the comparative advantage of technology to provide activities to each student that match his or her individual level of preparation.

[20] Current policy recommendations continue to stress the above-mentioned focus on student heterogeneity and personalization. For example, a recent World Bank country diagnostic for India refers to “remedial instruction” and “classroom practices that engage students at their current levels of learning” as “interventions that have shown promise in improving foundational learning in schools.” (The World Bank, 2018a, pg. 62).

1.4 Diagnostic assessment of skills

1.4.1 Sampling and sample

The study population consists of all fourth-, sixth-, and eighth-graders in 18 major Indian states and one union territory, who attended public schools between January and September 2009.[21] These states were selected for counting with more than one percent of India's total population. The study's sampling frame thus included 421 districts and their 657,787 government-run schools[22], with an enrolment of 25,519,296 students across these three grades-levels.[23]

To warrant representativeness, sampling followed a multi-stage, stratified cluster design, in which schools were selected using a probability-proportional-to-size (PPS) technique. In a first step, 48 districts were selected at random, stratifying by state and districts' level of development (using the Human Development Index or literacy rates, depending on data availability). Either two or four districts were selected by state; the number of districts and schools to be sampled per state followed sample size calculations that account for within-cluster variation, seeking to provide (overall) score estimates with a 0.1 standard deviation confidence band around the mean. In a second step, 2,399 schools were selected at random, determining a school's selection probability along with student enrollment numbers.[24]

Across the 2,399 schools, the effective study sample comprises of the 101,084 students who were present on the day of the assessment. This sample includes 29,513 fourth-, 35,604

[21] Our study sample and data rely on the Student Learning Study (SLS). SLS was conducted by Educational Initiatives, a leading Indian education firm, in collaboration with State Governments, with support by Google.org, and with additional advice from national and international expert bodies.

[22] In this context, with "government-run schools", we refer to "DOE" and "Local Body" schools. As of the 2015-16 school year, approximately 74 percent of Indian schools serving primary and upper-primary students are public schools (Mehta, 2017) and 65 percent of primary school students were served by public school (UNESCO Institute for Statistics, 2018).

[23] Accordingly, the study covers 71.7 percent of the total Indian government school population, in these grade-levels.

[24] See the technical report for the SLS study, for a more detailed description of the sampling strategy (Educational Initiatives, 2011).

sixth-, and 35,967 eighth-graders, of whom 51.2 percent are female. Approximately 67.4 percent of enrolled students took the study's mathematics assessment, reflecting similar net attendance and absentee rates reported elsewhere (IIPS, 2007; ASER, 2010).

Finally, one may be worried that our study's statistical model development fit the particular data at hand only, and not the larger population. To avoid such "overfitting", this paper uses a random split sample (RSS) technique. We assign a random 50 percent of sample students to a "training" data-set (used for model development) and the remaining students to a "holdout" data-set (used for estimation) (cf. Chen and de la Torre, 2014).[25] In determining these two subsamples, we stratify by state, grade-level, assessment language, and gender.[26]

1.4.2 Test characteristics

Student math skills were assessed with written, grade-level specific tests, whose items were scored as correct or incorrect. Tests were administered in 13 different languages and students were given 120 minutes to answer all questions on the test. Test development included three stages prior to administration, including curriculum and textbook reviews, small-scale pre-testing, and large-scale piloting with item analysis. A separate technical report (Educational Initiatives, 2011) provides detailed information on test development, including evaluator manuals, training materials, and an in-depth description of translation and harmonization procedures.[27]

The study's mathematics assessments were designed to reflect student's ability to apply their knowledge in specific content domains. Thus, items are mapped to specific abilities that are expected to be prerequisites to solve a given question correctly. In this

[25] Of course, alternative methods such as leave-one out, bootstrapping, or k-fold cross validation may be used as well. Given the large sample, we prefer the straightforward RSS strategy.

[26] We intend to divide each strata into two groups of equal sizes. Group sizes may differ slightly when the number of students in a stratum is uneven. In this case, we randomly assign these left-over observations (or "misfits") to one of the two groups, within each stratum (cf. Carril, 2017).

[27] The same report also provides copies of all exams and their items, as administered to students (in their English language version).

paper, we focus on content domains that were covered in all three grade-levels. This explicit skill-to-item mapping allows us to focus on five abilities, namely “Fractions and Decimals”, “Measurement”, “Number Concepts and Number Theory”, “Operations on Whole Numbers”, and “Shapes and Geometry”. Our study assesses ability levels across grade-levels and therefore, we drop items for skills that cannot be expected to be mastered in grades six or below—for example, we remove questions that were designed to capture a student’s mastery of algebra (we remove a total of 25 of such items from the original SLS instrument).[28]

A total of 91 different items map to the skill areas described above. 40 of these items were asked in grade four, 41 in grade 6, and 33 in grade 8. Note that our paper intends to measure mastery of skills on a common scale, requiring common “anchor” items both across grade-levels and across skills. 20 items are used to link across grade-levels. The assessment also allows for linking across any of the measured skills (as anchors are found in each of the skill-by-grade combinations). In addition, items that tap into skills of grade-levels beyond grade four were coded as such, using binary indicators (with 23 items reflecting grade-levels five or six, and 19 items reflecting grade-levels seven or eight.)[29] To give an example, Figure A2 provides three sample items measuring a child’s “whole number operations” skill, along with their respective grade-level.[30] Appendix Table A1 provides an overview of all test items. The same table also indicates the assessment’s skill-to-item mapping in its final, revised version (see section 1.4.4, below).

[28] In addition to these five domains, questions were categorized as being either “straightforward” or “not straightforward”. In direct reference to Bloom *et al.* (1984), with “non-straightforward items”, the test development team thus sought to tap into students’ ability to develop deeper understanding, going beyond factual or procedural knowledge. Future research may focus on students’ depth of understanding, within each domain.

[29] The remaining items reflect grades one through four. We follow Educational Initiatives’ mapping of test questions to grade levels.

[30] Here and elsewhere, item IDs refer to the question paper(s) and question order. “F” refers to the grade four, “S” to the grade six, and “E” to the grade eight test. For example, item F18S11 is an anchor item with question number 18 on the fourth-grade test and question number 11 on the sixth-grade test.

1.4.3 Methodology

This paper's psychometric analyses rely on a Cognitive Diagnosis Model.[31] 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). Accordingly, when we refer to a student's diagnosed "mastery" or "proficiency" profile, we thus refer to her empirical classification into a group of students who have mastered all assessed skills, none, or any combination thereof.

CDMs differ in terms of their complexity, as individual skills are related to students' propensity to answer individual test questions correctly. For example, does the model allow for multiple skills to interact? This paper relies on the generalized deterministic inputs, noisy and gate (G-DINA) model for dichotomous items (de la Torre, 2011), which subsumes other strategies that, for example, impose compensatory/disjunctive vs. non-compensatory/conjunctive relationships among attributes (Henson *et al.*, 2009).[32] Hence, we begin with the most general, most complex, modeling approach; in additional analyses, we thereafter explore whether this model can be simplified.

As common for CDMs, the G-DINA model (and the reduced models it subsumes) 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 mastery profile of each learner is described by a latent vector of dichotomous entries that each indicate whether an examinee has mastered

[31] CDMs are also referred to as "Diagnostic Classification Models". For a short, accessible introduction, see de la Torre *et al.* (2016). For a comprehensive introduction, see Rupp *et al.* (2010). For additional comparisons of CDMs with item-response theory-based (IRT-based) approaches, see Graf (2008) and Rupp and Templin (2008). For a more recent "didactically oriented" application of CDMs, see Jurich and Bradshaw (2014). For applied, technical guidelines and related software, see George and Robitzsch (2015). For an alternative, network-analytic strategy, see Pearlman (2011).

[32] A multitude of models have been developed, including approaches that model the underlying traits, or "attributes" as polytomous rather than dichotomous, and models that largely differ with respect to whether and how attributes are expected to interact as they determine a respondent's propensity to answer a given item correctly (see Rupp *et al.*, 2010, 158 et sqq.).

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. For example, for those items reflecting only the two attributes a_1 and a_2, the probability *P* of a correct response is modeled as follows.

$$P(X_j = 1|\boldsymbol{\alpha}^*_{lj}) = \lambda_j + \lambda_{j1}\alpha_{l1} + \lambda_{j2}\alpha_{l2} + \lambda_{j(1*2)}\alpha_{l1}\alpha_{l2} \tag{1.1}$$

Thus, in the above example, a student who has not mastered either skill would still be expected to answer the item correctly with a probability of λ_j (e.g., by guessing the correct answer). The probability of students with just one of the two required skills a_1 and a_2 would differ from this previous student's probability, as indicated by the respective main effects λ_{j1} or λ_{j2}. Finally, the probability for a master of both skills is indicated by the sum of all four elements in λ_j, including the interaction effect $\lambda_{j(1*2)}$.

More generally, following de la Torre (2011), a respondent's probability of solving an item can be expressed as

$$P(X_j = 1|\boldsymbol{\alpha}^*_{lj}) = \lambda_{j0} + \sum_{k=1}^{K^*_j} \lambda_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K^*_j} \sum_{k=1}^{K^*_j-1} \lambda_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \lambda_{j12\ldots K^*_j} \prod_{k=1}^{K^*_j} \alpha_{lk} \tag{1.2}$$

, where λ_{j0} reflects the probability of a correct answer to item *j* for non-masters (the intercept, or "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.

1.4.4 Psychometric analysis

Our psychometric analysis proceeds as follows. We begin our analyses with the training sample, estimate a saturated G-DINA model (using marginal maximum likelihood), and screen out items with limited information. Thereafter, we validate and refine the mapping of items to skills (i.e., the "Q-matrix"), following de la Torre and Chiu (2016), and through a qualitative review. Next, following Ma *et al.* (2016), we use the same training data-set to establish whether a more parsimonious, reduced model may be used without a significant loss in model data fit, for each of the test's items. We also compare (and rule out) that model fit could be improved by using a log-linear or logit link, instead of an identity link function. Lastly, after this model-selection procedure, we estimate our final cognitive diagnosis model on the holdout sample, discuss the model fit, and present polychoric correlations among the five skills. In the following subsections, we report on each of these steps in greater detail.

Item screening

We begin by estimating a preliminary G-DINA model, to identify items that provide limited information regarding a respondent's skill profile (using the training sample). First, students with differing skill levels should vary in their ability to solve a question correctly. A given item has low discriminatory power for a given attribute if a student who has mastered the attribute has the same (or similar) probability of solving the item as compared to a non-master. Second, students who are masters on attributes that are measured by an item should have a greater chance of answering said item correctly, as compared to non-masters (monotonicity). We remove five items that exhibit low discrimination or violate monotonicity. None of these five questions would have served as anchor items.[33]

Refined item-to-skill mapping

We then use our training sample to validate and refine the study's item-to-skill mapping empirically, following de la Torre and Chiu (2016). While this procedure is data-driven, we also investigated the resulting changes qualitatively, by presenting them to members of the test development team. A few observations stand out, corroborating the study's Q-matrix design. First, no changes were made to "Shapes and Geometry" and "Measurement", and only a single item was changed with respect to "Decimals and Fractions". This result coincides with the notion that respective skills can be rather easily attributed to items.

[33] The removed questions are items E11, E22, E45, E46, and E51.

Secondly, for five items we added the "Whole Number Operations" dimension. Almost all of these cases (four), reflect a shift from number sense to the whole number operations skill. Figure A2 in the Appendix provides a sample item that exemplifies this change. We proceed with the suggested Q-matrix modification (see Appendix Table A1 for the final item-to-skill mapping). Altogether, we only modify the item-to-skill mapping for six items, which may reflect the assessment's careful instrument design, along with the same dimensions we investigate in this research.

Accounting for grade-level expectations

So far, our analyses of student mastery have assumed a single scale per skill. This approach is reasonable for the above item screening and refinements of the item-to-skill mapping; however, substantively, little can be learned from these results that indicate whether fourth-graders or sixth-graders have mastered an attribute whose scale comprises eighth-grade material. In contrast, we deem it more useful to assess whether fourth-, sixth-, and eighth-graders have mastered a basic, fourth-grade level understanding for each of the five skills. In line with probabilistic Guttman models of ordered latent traits (Proctor, 1970), the following analyses therefore account for whether items measure either fourth-grade material only, or an additional level of mastery that also requires knowledge of more advanced material (i.e., grades five to eight).

We account for grade-level expectations by modifying our psychometric methodology in three ways. First, we extend the Q-matrix to ten columns, indicating whether each item captures an attribute either not at all, at a fourth-grade level, or both at a fourth-grade level and at levels beyond fourth grade. Note that this approach perceives of the fourth-grade level as a prerequisite of more advanced mastery; it is not possible to have reached an eighth-grade level of understanding without "graduating" from the previous level. Our second modification accounts for this hierarchical attribute structure by constraining the percentage of students in these attribute categories to zero. Note also that our strategy focuses on the *additive* nature of (prerequisite and) advanced skills; our conceptualization of advanced mastery jointly captures the main effect of a more advanced level and its interaction with a fourth-grade level. For items measuring a single attribute, our third change to the model therefore constrains the G-DINA to an additive cognitive diagnostic model (A-CDM). Effectively, this strategy drops all interactions from Equation 1.2; the A-CDM item response function is given by

$$P(X_j = 1|\boldsymbol{\alpha}^*_{lj}) = \lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk}\alpha_{lk} \tag{1.3}$$

Model parsimony

For those items measuring more than one skill, we considered constraining our initial G-DINA estimation to a more parsimonious model.[34] We allow for four competing models with varying assumptions regarding the (either compensatory or conjunctive) relationships among skills: the DINA (deterministic input, noisy, and gate), DINO (deterministic input, noisy, or gate), A-CDM (additive cognitive diagnostic model), and RRUM (reduced reparametrized unified model).[35] For model comparison at the item-level, we rely on two-step likelihood ratio tests as proposed by Sorrel *et al.* (2017). Moreover, we calculate the dissimilarity index as suggested by Ma *et al.* (2016) to compare the G-DINA with the remaining four, simpler CDMs. From these analyses, we conclude that a more parsimonious model should not be used, for these items.

Item characteristics

Appendix Table A1 in the Appendix provides item parameter estimates from the final model, as developed above, and estimated on the training sample. These parameters are held fixed as we re-estimate the model on the holdout sample. Appendix Figure A3 presents respective "item characteristic bar charts" (ICBCs).[36] These ICBCs depict a student's

[34] For research that takes a similar approach, see Tu *et al.* (2017).

[35] See Li *et al.* (2016) for a concise description of these models, in direct comparison to G-DINA.

[36] This terminology follows Bradshaw *et al.* (2014).

expected probability of solving an item correctly, conditional on her level of mastery for the skills measured by the given item.[37] For simplicity, if items measure multiple skills or grade-level expectations, we only display the expected probability for students who have not mastered any, versus students who have mastered all measured attributes. See Appendix Table A1 to calculate probabilities of students with mastery of select attributes.

On average, the items' intercept is 0.27, potentially reflecting that items can be solved by guessing, even if students have not mastered (any of) the respective skill(s). We moreover find that items are of diagnostic utility in terms of their ability to discriminate masters from non-masters. On average, masters of all measured attributes have a 46 percentage point (pp) higher probability of solving a given item correctly, in comparison to the aforementioned students without any mastery.[38] Given the large sample size, all item parameters are measured with very tight error bands (the largest standard error is 0.0028).

Fit and attribute reliability

Item fit is found to be good when the model is estimated on the holdout sample, as indicated by an average root-mean squared deviation (RMSD) item fit statistic of 0.055 (and only few items exhibiting an RMSD statistic above 0.1).[39] We moreover report on four, common measures of absolute model fit (cf. Chen *et al.*, 2013). We conclude that the model fits the holdout sample's data well, given the following fit statistics: a mean of absolute deviations in observed and expected correlations of 0.054, a standardized mean square root of squared residuals of 0.070, a mean of absolute deviations of residual covariances of 0.011, and a mean of absolute values of the centered Q_3 statistic of 0.066.

In terms of reliability, we find that the test's classification consistency is moderate, at the individual level. Following Cui *et al.* (2012), our calculations suggest that the assessment's

[37] Readers familiar with item response-theory (IRT) may draw parallels between ICBCs and IRT-based item characteristic curves (ICCs).

[38] This comparison (or "discrimination index") follows de la Torre (2008).

[39] RMSD is also sometimes interchangeably referred to as the "root mean square error of approximation" (RMSEA). See Kunina-Habenicht *et al.* (2009).

consistency ranges from 0.55 to 0.99, per attribute, for an overall consistency of 0.66. This finding is less problematic for the present study as its stated goal is to report on *aggregate* mastery levels, by gender, grade-level, and location (see below). However, we caution from alternative uses of the same instrument for purposes that require the classification of individual students (e.g., as a means to provide targeted remediation). In addition, we find low levels of classification accuracy (of 0.38, on average), i.e. a high likelihood that an individual test-taker's classification does not agree with her true latent class. Again, this is less problematic given our focus on aggregates. Moreover, as expected, we observe higher levels of classification accuracy for the prerequisite, fourth-grade level of skills (of 0.59, on average), favoring our focus on this level of understanding.

Correlations between skills

Table 1.1 presents polychoric correlations among the five skills (at the fourth-grade level of mastery). We highlight three observations. First, correlations are high (0.90, on average), but lower than in other, retrofitted analyses whose results either do not seem to detect distinct traits or analyze attributes which are not distinguishable ontologically (for a discussion, see Bradshaw *et al.*, 2014). Secondly, the reported correlations support the face validity of the test—for example, as may be expected from learning theories (Clements and Sarama, 2014), skills such as geometry and whole number operations are correlated less strongly, as compared to number sense and whole number operations. Lastly, "Number Concepts" shows the greatest correlation with other attributes, supporting the notion that this skill may serve as a foundation for the remaining four areas.

1.4.5 Excursus: Limitations of "percent correct" as an alternative analytic strategy

In Section 1.2, we referred to other research that groups items into sub-sets, along with sub-skills and grade-levels. This approach thereafter reports on the percentage of students who correctly solved a given sub-set of items (e.g., the percentage of fourth-graders who

Table 1.1: *Polychoric correlations between skills*

	Fractions and Decimals	Measurement	Number Concepts	Operations on Whole Numbers	Shapes and Geometry
Fractions and Decimals	1				
Measurement	0.911	1			
Number Concepts	0.896	0.916	1		
Whole Number Operations	0.869	0.914	0.943	1	
Shapes and Geometry	0.903	0.845	0.915	0.862	1

Notes. This table reports on polychoric correlations between skills.

solved fourth-grade geometry items correctly). How does this "percent-correct" strategy compare to our choice of a Cognitive Diagnostic Model (CDM)?[40]

In Table 1.2, we present results from this alternative analytic strategy. This table reports on the percentage of students answering select fourth-grade items correctly, by sub-skill and students' enrolled grade-level.[41] We distinguish the set of all fourth-grade items administered to students of a given grade-level ("any grade-four items") from sub-sets of fourth-grade items that allow for comparisons across grade-levels (as they have also been administered to peers in other grades). Several limitations become apparent, severely restricting our ability to measure a common fourth-grade mastery-level for students across grades, with a percent-correct strategy.

To begin, Table 1.2 shows how the overall number of grade-four questions administered is low. For higher grades, test developers naturally seek to cover materials at grade-level, when they construct a test. Accordingly, we document only between one and three fourth-grade questions, per sub-skill, that have been administered to eighth-grade students. In addition, test developers may also include questions that tap into skills from earlier grades (i.e., grades one to three). For the given test presented to fourth-graders, for example, 15 of the 40 items are mapped to materials from grades two and three.

[40] Readers familiar with comparisons of percent-correct approaches and item-response theory (IRT) may skip this section—we highlight similar benefits, for CDMs. At the same time, we cannot deploy an IRT-based strategy per sub-skill, given the low number of items per sub-skill.

[41] We use the same assessment and the same "holdout" sample of students we use elsewhere, to estimate and present CDM-based results.

Table 1.2: *Limits of percent correct as a measure of fourth-grade sub-skills, across grade-levels*

	Fractions and Decimals	Measurement	Number Concepts	Operations on Whole Numbers	Shapes and Geometry
Fourth-graders					
Any grade-four items	49.0% (5)	27.0% (3)	62.1% (6)	54.6% (10)	43.8% (2)
Fourth-graders only	61.6% (3)	26.8% (1)	61.3% (5)	58.5% (6)	42.1% (1)
Fourth- and sixth-graders	29.7% (2)	22.7% (1)	n/a (0)	48.9% (4)	45.4% (1)
Fourth- and eighth-graders	39.9% (1)	26.8% (2)	63.5% (1)	n/a (0)	n/a (0)
Administered to all students	39.9% (1)	22.7% (1)	n/a (0)	n/a (0)	n/a (0)
Sixth-graders					
Any grade-four items	28.5% (3)	21.1% (1)	38.4% (1)	43.5% (8)	46.7% (3)
Fourth- and sixth-graders	35.7% (2)	21.1% (1)	n/a (0)	61.5% (4)	55.6% (1)
Sixth-graders only	15.2% (1)	n/a (0)	n/a (0)	28.5% (4)	48.0% (1)
Sixth- and eighth-graders	46.3% (1)	21.1% (1)	38.4% (1)	n/a (0)	33.1% (1)
Administered to all students	46.3% (1)	21.1% (1)	n/a (0)	n/a (0)	n/a (0)
Eighth-graders					
Any grade-four items	50.1% (1)	41.2% (2)	70.8% (3)	79.7% (1)	45.2% (1)
Fourth- and eighth-graders	50.1% (1)	41.2% (2)	83.6% (1)	n/a (0)	n/a (0)
Sixth- and eighth-graders	50.1% (1)	34.4% (1)	38.9% (1)	n/a (0)	45.2% (1)
Eighth-graders only	n/a (0)	n/a (0)	90.1% (1)	79.7% (1)	n/a (0)
Administered to all students	50.1% (1)	34.4% (1)	n/a (0)	n/a (0)	n/a (0)

Notes. This table reports on the percentage of students answering select fourth-grade items correctly, by sub-skill and students' enrolled grade-level. We distinguish the set of all fourth-grade items administered to students of a given grade-level ("any grade-four items") from sub-sets of items that have also been administered to peers in other grades. Number of items in parentheses. Items mapped to more than one sub-skill are included multiple times.

Table 1.2 also highlights how the number of questions that are comparable across grade-levels is low. For two of the skills, there is only one item that was administered to students across all three grades (for Fractions and Decimals, Measurement). For the remaining three skills (Number Concepts, Operations on Whole Numbers, and Shapes and Geometry), there is no comparable item. Even if only two grade-levels are compared with each other, this only improves slightly. For example, there are at most two items that allow for a comparison of fourth-graders with eighth-graders.

These results also point to another limitation: they are heavily influenced by the inclusion (or exclusion) of individual questions. For instance, in our sample, fourth-graders solved 49.0 percent of fourth-grade questions measuring "Fractions and Decimals". This number is 61.6 percent for those fourth-grade questions only administered to them, 29.7 percent for the set of fourth-grade questions that allow for comparisons with sixth-graders, and 39.9 percent for the fourth-grade item that allows for comparisons across all three grades.

We mention three additional shortcomings only briefly. Some items measure multiple sub-skills and, hence, it remains unclear how to interpret their results. Percent correct also does not specify "mastery"—a concept that appears highly policy-relevant. Lastly, percent correct usually assumes that all questions should be weighted equally—the strategy does not account for the difficulty of questions, for the ability of questions to discriminate low-ability from high-ability students, or for students who may have simply guessed a question's answer correctly.

In contrast, CDMs leverage common items as "anchors", include *all* test questions on a test, and estimate student ability on a common (here: fourth-grade) scale ("vertical linking").[42] They also provide an empirical estimate of student mastery. They moreover account for questions that tap into multiple sub-skills (and their potential interaction). Finally, CDMs explicitly model item difficulty, their ability to discriminate, and students' ability to guess a correct answer.

1.4.6 Results

This section reports on our study's results, presenting aggregate skill profiles of Indian students by grade-level, location, and gender. We begin our presentation of results by focusing on overall levels of mastery by students' (enrolled) grade-level, for each of the five skills. In doing so, we contrast whether the measured skill domains are covered by common assessments of the Global Learning Crisis, or not. We then continue with discussions of geographic heterogeneity. The last set of results assesses differences in mastery levels across male and female students.

Main results

Table 1.3 provides aggregate mastery levels at fourth-grade proficiency, by students' (enrolled) grade-level and skill. Three observations stand out from this table. First, learning

[42] CDMs only "lose" items to the extent that students in higher grades are tested on additional sub-domains (such as algebra, in our case.)

levels are low across the five areas—for example, only about 59 percent of eighth-graders have mastered a fourth-grade proficiency of "Fractions and Decimals" and "Shapes and Geometry", respectively; only 57 percent of them have mastered the "Measurement" domain.[43] Secondly, within each grade, we observe large differences across skills. For instance, 65 percent of fourth-graders have mastered "Number Concepts", whereas only 43 percent of them exhibit mastery in "Measurement". Thirdly, across the five skill areas, we find large differences as to whether students "catch up", by mastering fourth-grade proficiency at a later grade-level.[44] In particular, the percentage of masters of number concepts and whole number operations improves, reaching 75 and 71 percent in eighth-grade, respectively. In contrast, when comparing fourth-, sixth-, and eighth-graders, there is little difference in the share of students who have mastered "Fractions and Decimals" or "Shapes and Geometry".

Table 1.3: *Mastery of fourth-grade skills, by students' enrolled grade-level*

Class	Fractions and Decimals	Measurement	Number Concepts	Operations on Whole Numbers	Shapes and Geometry
4	0.515	0.429	0.646	0.485	0.535
6	0.533	0.489	0.664	0.593	0.560
8	0.587	0.571	0.748	0.713	0.587

Notes. This table reports on the probability that a student has mastered fourth-grade material, by students' enrolled grade-level and skill.

How does a focus on Number Concepts and Whole Number Operations—as done by the ASER test, for example—affect assessments of the learning crisis? We address this question through Figure 1.1. With this figure, we complement Table 1.3 by showing the percentage of students who have mastered both of these skill domains (at a fourth-grade level), the percentage of students who have mastered the remaining three skill domains (again, at a fourth-grade level), and the percentage of students who have mastered all of these domains.

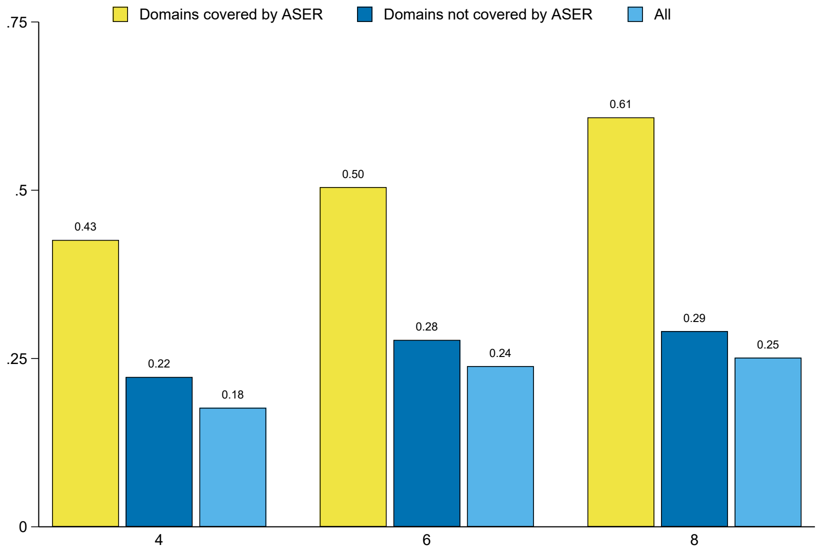
We observe how, across the three grade-levels, proficiency in Fractions and Decimals,

[43] We present unweighted results. An alternative approach may weight the percentage of masters by states' population size.

[44] Recall that we do not observe the same students over time. Strictly, we therefore cannot distinguish a "learning" effect from "cohort" or "compositional" effects. However, it appears unlikely that cohort or compositional effects would impact a subset of skills, only.

Measurement, and Shapes and Geometry remains far below the proficiency levels of the domains covered by ASER-type assessments (by 21, 22, and 32 percentage points in grades four, six, and eight, respectively). Moreover, proficiency increases for Number concepts and Whole Number Operations in the higher grades (to 50 and 61 percent among sixth- and eighth-graders, respectively). For the other three skills, however, these rates remain at their low levels (at 28 and 29 percent among sixth- and eighth-graders, respectively). Finally, we find that only a quarter of eighth-graders have mastered all of fourth-grade material—just slightly above the low proficiency-levels among their grade-six (24 percent) and grade-four (18 percent) peers. A sole focus on Number Concepts and Whole Number Operations would thus grossly overstate learning levels, and therefore strongly underestimates the extent of the learning crisis.

Figure 1.1: *Mastery of fourth-grade skills, by students' enrolled grade-level and skill domains*

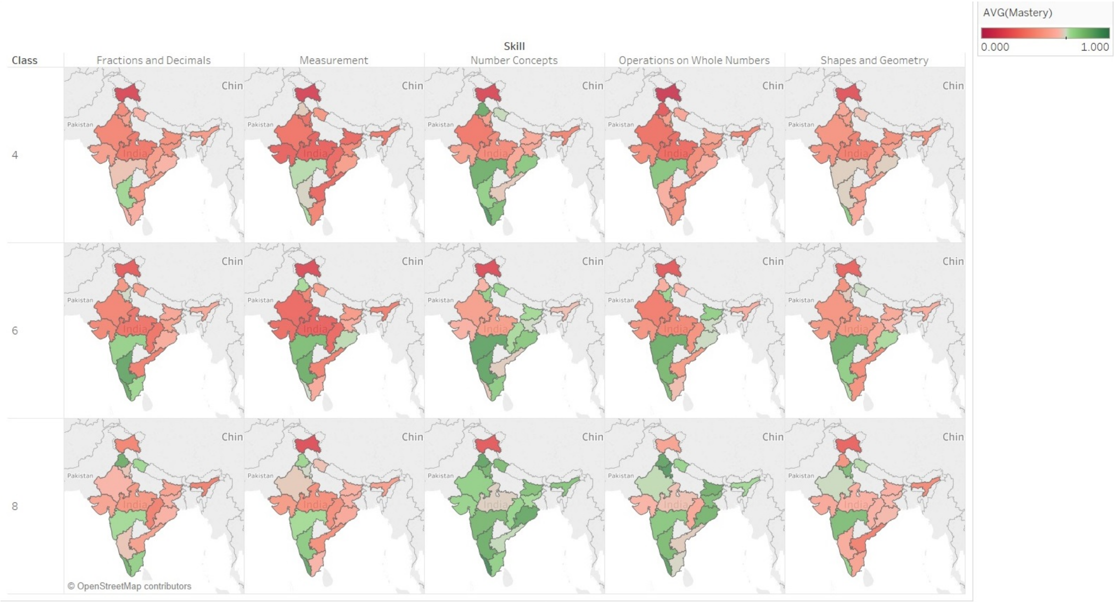


Notes. This figure reports on the probability that a student has mastered fourth-grade material, by students' enrolled grade-level and skill. Domains covered by ASER refers to Number Concepts and Whole Number Operations. Domains not covered refers to Fractions and Decimals, Measurement, and Shapes and Geometry.

Heterogeneous results by state

Our findings also point to large geographic heterogeneity in student mastery, across India's states. Figure 1.2 displays our visualization of aggregate mastery levels by (fourth-grade) skill, students' enrolled grade level, and states (see Appendix Table A2 for the figure's underlying estimates).[45] To take "Number Concepts" as an example skill, fourth-, six- and eighth graders in Jammu and Kashmir show the lowest share of masters (15, 17, and 23 percent, respectively); thus, eighth-graders rank below their fourth-grade peers from all other Indian states in the study. In contrast, we find that 91 percent of Kerala's sixth-graders have mastered this skill; a higher percentage than for eighth-graders in almost any other state (with the exception of eighth-graders in Chandigarh). Figure 1.2 moreover suggests that these geographic differences in student performance are relatively stable, both across grade-levels and across skills.

Figure 1.2: *Geographic heterogeneity in student mastery, by students' enrolled grade-level and skill domains*



Notes. This figure reports on the probability that a student has mastered fourth-grade material, by students' enrolled grade-level, skill, and state. In Delhi, only fourth-graders were tested.

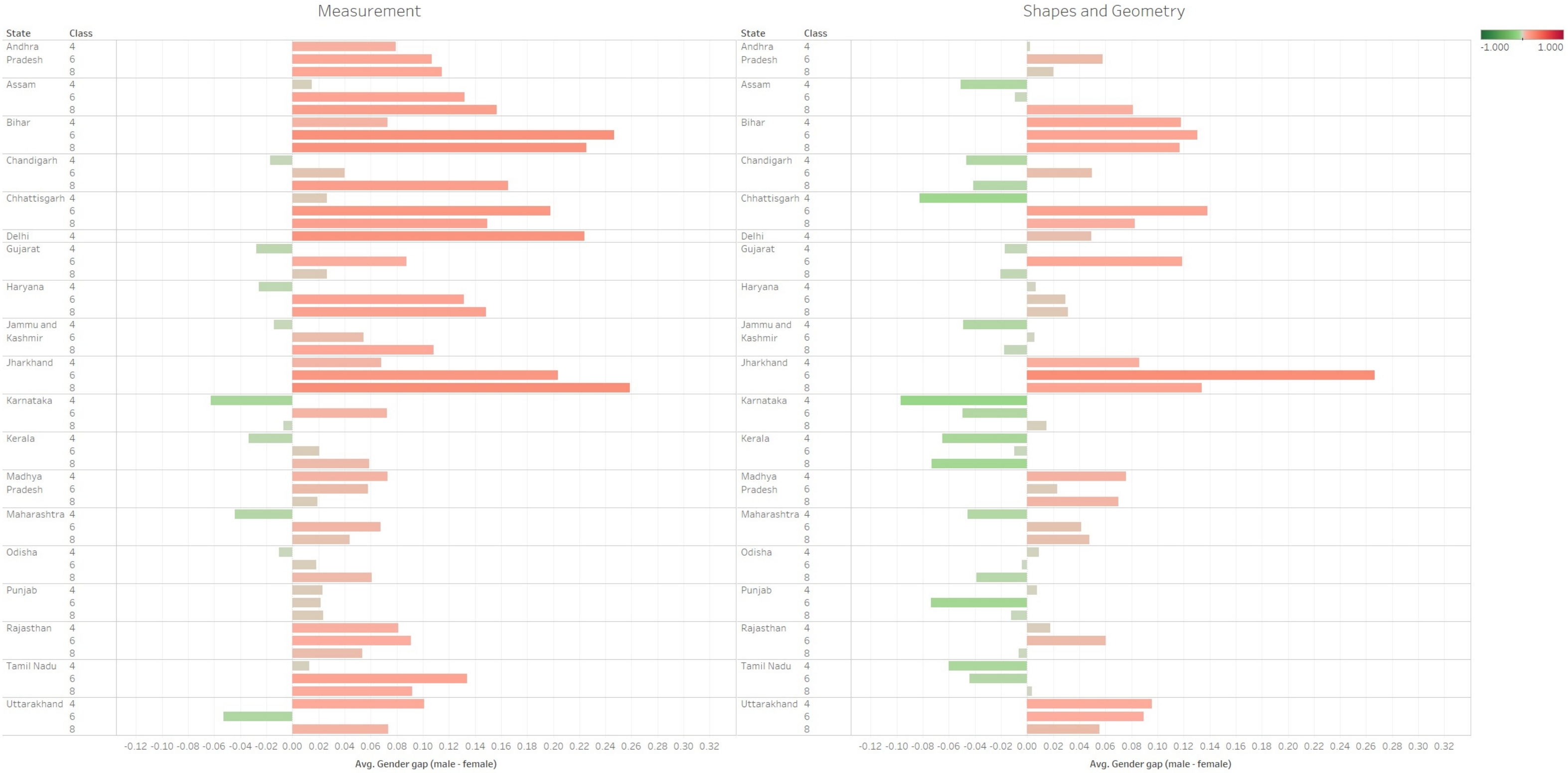
[45] We set the color scheme's midpoint (between red and green) to 66 percent. This choice is arbitrary. We do not make a normative statement and do not seek to imply that one third of non-proficient students is desirable (especially as mastery is measured on a common fourth-grade scale, including for sixth- and eighth-graders).

Heterogeneous results by gender

To investigate gender gaps, we furthermore calculate the difference in average mastery across male and female students. Overall, we find substantive differences in the prevalence of gaps across skills, from geometry (2.6 pp), to fractions (3.3 pp), whole number operations (4.9 pp), number sense (6.8 pp), and measurement (7.5 pp). Figure 1.3 presents more fine-grained results, by grade, state, and skill. Bars in red highlight if the percentage of masters is higher among boys, and bars in green highlight the opposite case. For simplicity, Figure 1.3 only presents results for the skill with the smallest gender gap (Shapes and Geometry) and for the skill with the largest gender gap (Measurement). See Appendix Table A3 for the figure's underlying estimates, including for the remaining three skills.

These more detailed results point to geographic heterogeneity in the prevalence of gender gaps. While gaps predominantly favor male students over female students, this relationship is particularly stark in Jharkhand and less pronounced in Karnataka and Punjab, for example. We also find notable differences across grade-levels. In Assam and Tamil Nadu, for instance, we observe how skill gaps only favor males in grades six and eight. Lastly, even though we found skills to be highly correlated (see sub-section 1.4.4 above), we also observe individual states in which gender gaps materialize more strongly for a particular skill (e.g. for Measurement, in Haryana or in Jammu and Kashmir).

Figure 1.3: *Gender skill gaps, by students' enrolled grade level and state*



Notes. "Gender skill gap" refers to the probability of male students who have mastered a skill, minus the respective probability for female students.

1.5 Conclusion

In this article, we reported on a cognitive diagnosis of students' mastery of mathematical skills, using large-scale, representative data from India. After connecting our study to previous research on learning levels and profiles in the developing world, we provided a short introduction to the study setting. In doing so, we noted a heightened interest in large-scale student assessments, yet a lack of measures that allow for the estimation of student skill profiles. While interest in student assessments has been growing over the past decade, many of these measures do not go beyond presentations of summary scores and the discussion of student performance on sample items. We also noted how several large policy initiatives share a common need for a more fine-grained understanding of students' mastery of skills.

This article reacts to this insight, with three main contributions. First, our study documents that detailed assessments of student mastery are in fact feasible; we show that more detailed information could be gained from current efforts to measure students' learning levels. Second, our research highlights that more fine-grained assessments and the reporting of disaggregated results are necessary. Such efforts could be spared if more parsimonious proxy indicators led to similar conclusions—yet, to the contrary, we find how less nuanced indicators, of a more limited skill domain, lead to a gross overestimation of learning levels. Third, we are confident that the study's results themselves, including their breakdown by geography and student gender, provide a first step to a more detailed understanding of what students can (and cannot) do, informing practitioners and policy makers.

We conclude by highlighting three avenues for future research. On the one hand, we provided evidence for the lack of reliability and accuracy of estimates, at the student level. This observation is less concerning for our presentation of aggregate mastery levels; however, this result calls into question whether the study's assessment could have been used to inform targeted decisions for individual learners. Secondly, the study's instrument development made a deliberate effort to measure the *depth* of student knowledge, through dedicated test items. Our research clearly benefits from this feature as each measure of skills captures learning beyond “the surface”, rote memorization, or the “simple” internalization of procedures. Yet, additional work may model more explicitly whether a deep state of mastery has been reached. Lastly, as our finding underlines the severity of the Global Learning Crisis, more research is needed to understand its underlying causes. For example, additional work may focus on misconceptions and the role of errors, as a means to address students’ low levels of learning. We strongly encourage additional research in these areas.

Chapter 2

Evaluating Teacher Evaluation – Evidence from Chile[1]

2.1 Introduction

Performance evaluations are among the most controversial attempts to improve teacher effectiveness, especially if they are based on student test score gains and if they are used to inform teacher compensation and dismissal.[2] In contrast, there are frequent demands for more comprehensive, "formative" evaluation systems (see Grissom and Youngs, 2016). In its call for such a formative approach, for example, the United States' largest teacher's labor union defines the core purpose of teacher assessment and evaluation as to "strengthen the knowledge, skills, dispositions, and classroom practices of professional educators" (as opposed to a "rewards-and-punishment framework") (National Education Association, 2019, 1).

Proponents of such formative evaluations often refer to Chile's national evaluation system as a best-practice example. Chile's system embraces core principles of a formative teacher

[1] Single-authored.

[2] For recent reviews of such evaluation policies, see Jackson and Cowan (2018) and Lovison and Taylor (2018).

evaluation approach, such as the promotion of professional development, open collaboration and transparency, the use of multiple measures, validation, links to clear teaching standards (based on a "Framework of Good Teaching"), and the system's co-creation with teachers. A recent World Bank report therefore concludes that, while "[p]utting in place a sound system of teacher evaluation is expensive and institutionally challenging", "Chile's comprehensive teacher evaluation system, Docentemas [sic], has shown that it can be done" (Bruns and Luque, 2014, 215 et sq.).[3]

This study evaluates the causal effects of repeat, formative performance evaluations—under Chile's national teacher evaluation system "Docentemás"[4]. The article answers three main questions. First and foremost, do these formative evaluations lead to increased teacher effectiveness, as measured by student learning? Second, how are potential mechanisms affected that are expected to enhance student learning? Formative evaluations seek to improve instructional practices and to alter commonly held beliefs among teachers—I therefore investigate impacts on these intermediary factors. Third, do evaluations affect less-experienced teachers more strongly? Previous research on the returns to teacher experience and on teachers' dynamic skill development suggests that productivity improvements predominantly occur during the first five to ten years on the job.[5] Hence, I assess heterogeneous effects of evaluations by teachers' level of work experience.

A key challenge for the estimation of causal effects of teacher evaluations is the endogeneity of a teacher's assignment to evaluations.[6] To overcome this challenge, this study's analytical strategy relies on a difference-in-difference estimation strategy. More specifically, I exploit a policy change in the assignment mechanism. In 2011, Chile passed a new law,

[3] The same report concludes that Chile's teacher evaluation system "remains the [Latin American] region's best practice example to date" (*ibid.*, 35).

[4] Formally, Docentemás is called the "Sistema de Evaluación del Desempeño Profesional Docente". Commonly, it is also referred to as "Evaluación Docente".

[5] For a recent overview, see Kraft *et al.* (2018b).

[6] The direction of this bias is unclear. Formative evaluations may be assigned to weaker teacher of greatest need of personal development. Yet, more motivated, stronger teachers may also self-select into formative evaluations as they seek out personal development opportunities.

requiring teachers ranked in the "basic" (the second lowest) performance category to be re-evaluated after two years (instead of four). My identification strategy leverages this variation across time. I show that, although the law is imperfectly observed, it sharply increased a "basic" teacher's likelihood of being newly evaluated after two years. In additional analyses, the article moreover confirms that common assumptions of a difference-in-difference estimator are met, and that its analyses are not compromised by systematic student sorting or by differential attrition.

These analyses rest on data sources with unusually comprehensive coverage of a national education system. For the years 2005 to 2015, I use teacher-classroom links to match data on the universe of elementary teachers in Chile's public schools, all teacher evaluations conducted in these years, administrative records for the universe of Chilean students (covering more than 30 million student-by-year observations), and results on standardized test scores for all Chilean fourth graders in mathematics and language. I further complement these data with information on teaching and teacher behaviors from teacher surveys, student surveys, and parent surveys.

The study's main results suggest that student learning remains unaffected by a teacher's requirement to undergo a formative evaluation, both in the year of the evaluation and in the year thereafter. Intent-to-treat effects are precisely estimated, ruling out positive impacts of 0.03 and 0.08 standard deviations, respectively. In analyses of potential mechanisms, the study documents how teacher beliefs and teaching behaviors also remain unimproved. If anything, the findings point to detrimental effects of formative teacher evaluations. These results do not differ for teachers with fewer years of work experience. In robustness checks, I moreover show how they do not depend on model specifications and that their qualitative conclusions remain unaltered if a slightly modified assignment definition is used.

These analyses and their findings are novel in three distinct ways. They represent only the second causal investigation on the impact of formative teacher evaluations on student learning (and the first to include additional analyses of potential mediators). They moreover offer first evidence on the impact of repeat evaluations (i.e., the effect of *regularly* subjecting teachers to evaluations). To my best knowledge, this study furthermore provides the first quasi-experimental assessment of a national workforce evaluation system's effects on worker productivity—hence, it may also offer interesting insights for public human resource management beyond the education sector.

The study thus contributes to several strands of a nascent literature within the economics of education and public economics, on the effectiveness of formative performance evaluations. One related body of literature has studied the effects of teacher evaluations on other teacher-level outcomes, such as professional improvement activities (Koedel *et al.*, 2019), effort (Aucejo *et al.*, 2019), job satisfaction (Koedel *et al.*, 2017), and labor market responses (Sartain and Steinberg, 2016). Another collection of studies has focused on the effect of sub-components that may be part of formative evaluation systems, including in-person classroom observations (Burgess *et al.*, 2019; Kane *et al.*, 2019), peer collaboration (such as lesson study and instructional rounds) (Gersten *et al.*, 2010; Louis and Marks, 1998), tutoring, mentoring, and coaching (Allen *et al.*, 2011; Papay *et al.*, 2019; Kraft *et al.*, 2018a; Kraft and Hill, 2019), the provision of formative feedback (Garet *et al.*, 2017), and the release of teacher performance scores (Bergman and Hill, 2018; Pope, 2019).

Strikingly, however, there is so far only one other study on the causal effects of formative teacher evaluations on student performance.[7] For a sample of 105 mid-career teachers in Cincinnati Public Schools, Taylor and Tyler (2012) apply a teacher fixed-effects strategy. They find that math test scores of students whose teacher was evaluated in the previous year increased by about ten percent of a standard deviation. Taylor and Tyler (2012) also conclude that this effect is greater for teachers whose previous performance was lower. They cannot reject the absence of effects on reading scores.

The remainder of this paper proceeds as follows. The next section briefly describes theoretical considerations. Section 2.3 gives a short introduction to Chile's teacher evaluation system, and Section 2.4 introduces the proposed estimation framework. This is followed by

[7] Steinberg and Sartain (2015) study the effects of a formative teacher evaluation pilot program on *school* performance. In a randomized trial with 92 primary schools, they find positive effects on language, after one year, but no statistically significant effects on math.

a description of the paper's data sources, in Section 2.5. Subsequently, Section 2.6 provides summary statistics for the analytic sample and scrutinizes the study's internal validity. Section 2.7 presents results and Section 2.8 concludes.

2.2 Theoretical considerations

From a theoretical viewpoint, there are no clear expectations considering whether, if at all, teacher evaluations have a positive or negative impact on teacher effectiveness. A short review of five theoretical lenses illustrates this point. To begin, as discussed by Papay (2012) and Taylor and Tyler (2012), a *human resource development view* predicts increased teacher performance and improvements in student learning, as evaluations provide teachers with information on how to improve. This approach also suggests that evaluations allow teachers to learn about skill and performance expectations. Further, a *professionalization argument* hypothesizes that student learning may be improved if increases in teacher evaluations allow teaching to graduate from a "second grade" to a "full" profession (Johnson and Fiarman, 2012; Mehta, 2013).[8] Next, *principal agent theory* may also predict positive effects through improved information on effort, such as an increase in the ability of principals and parents to monitor effort and performance (Hölmstrom, 1979; Milgrom and Roberts, 1992). At the same time, *the multi-tasking model* (*cf.* Jacob, 2005) posits that, while evaluations may increase teachers' efforts to improve on those tasks that are observed by the performance measure, teachers may shift their efforts away from other, unobserved tasks. Thus, under this model, ambiguous effects can be expected. Finally, critics of teacher evaluations point to rather practical concerns and to an *opportunity-cost argument*, suggesting that teacher evaluations may take up scarce financial resources and teachers' work-time (*cf.* Taut *et al.*, 2011).

[8] Mehta (2013) argues that teacher evaluations may also hinder professionalization if they are used to promote external teacher accountability based on test-scores; he promotes approaches instead.

2.3 Teacher evaluation in Chile

This section provides a short overview of Chile's teacher evaluation system, Docentemás, focusing on those characteristics that inform the study's identification strategy.[9] Docentemás was introduced in 2003 as a standards-based, formative assessment system that is tied to the country's national Framework of Good Teaching ("Marco para la Buena Enseñanza").[10] In 2005, participation became mandatory, for all public schools in the country.[11]

A teacher's evaluation includes four components with differing weights, as follows: A self-evaluation (10%), a third-party reference report (10%), a peer evaluator interview (20%), and a teacher performance portfolio (60%).[12] The latter consists of a teacher's submission of a portfolio describing an eight-hour learning unit and of an announced video recording of a class. Sub-scores for each of these components are aggregated to a single, continuous performance score, which is then used to rate teachers along four performance levels: unsatisfactory, basic, competent, and outstanding. The continuous score ranges from 1 to 4, and values of 2, 2.5, and 3 are used as cut-scores, respectively. However, a teacher's rating may be modified by a Municipal Evaluation Commission before it becomes final (modifications occur in approximately five percent of cases).

The overall evaluation spans one year. Its process begins with a teacher's nomination in April and continues with the submission of portfolios, recordings, self- and peer-evaluations between August and October, as well as with the third-party report in November. Grading takes place in December and January, final grades are decided upon in February and March,

[9] See a recent OECD review for a comprehensive presentation, including information on Chile's school system, in English (Santiago *et al.*, 2013). See Manzi *et al.* (2011) for a detailed presentation of Chile's teacher evaluation system, in Spanish.

[10] The Framework is based on Danielson's Framework of Good Teaching and the Measures of Effective Teaching (MET) Project (see Santiago *et al.*, 2013).

[11] For 2010, Manzi *et al.* (2011, 26) report that 96% of all public teachers complied with their legal obligation to participate in the evaluation. I calculate that, in 2015, 80% of eligible teachers had been evaluated at least once. This calculation focuses on elementary teachers in public schools, teaching either mathematics, reading, or "general".

[12] These weights change in the case of follow-up evaluations after a rating in the bottom category. The adjusted weights are as follows: Self-evaluation (5%); third-party reference report (5%); peer evaluator interview (10%); teacher performance portfolio (80%).

teachers receive their results in March with detailed written feedback, and further reports are distributed to other parties in April.[13] Interestingly, results for the largest evaluation component (centralized, anonymous ratings of the teacher performance portfolio) thus only become available *after* the remaining three components have been scored.[14]

In 2011, a new law (Ley 20.501) introduced changes to the consequences of a teacher's performance rating. Generally, public teachers are required to be evaluated at least once every four years.[15] Yet, under the new law, teachers rated as "basic" must be re-evaluated after two years.[16] The law came into effect in 2011, but it did not affect teachers retroactively, based on their previous performance ratings. Below (in Section 2.7.1), I show that the policy sharply increased a "basic" teacher's probability of getting re-evaluated after two years. In contrast, teachers in the pre-policy period and teachers with a higher rating were not re-evaluated.

For "basic" teachers, the new law did not result in other changes—whether with respect to their job security, their access to incentive schemes, or their professional development, for example. In terms of teacher turn-over, since 2011, a teacher with a "basic" rating must leave the system if her rating does not improve in the next two assessments. However, the potential reduction in a "basic" teacher's job security only applies after her *second* follow-up evaluation. Since 2011, some principals are also allowed to dismiss up to five percent of "basic" *and* "unsatisfactory" teaching staff. Yet, this change only applies to a subset of principals who have been hired through a competitive process. Further below, I investigate—and do not find support for—the law's impact on "basic" teachers' turn-over.

[13] The Chilean school year begins in March and ends in December.

[14] Arguably, this feature reduces the likelihood of score manipulation around the three cut-scores—I discuss this matter and its implications for the paper's econometric strategy further below.

[15] Docentemás covers all teachers in municipal schools above a set workload threshold. Teachers are nominated for evaluation by the head of their respective municipal school authorities ("Municipal Education Administration Department" or "Municipal Education Corporation"). New hires are not evaluated in their first year of service. Since 2006, teachers may opt out in their last three years before qualifying for retirement.

[16] Teachers with an "unsatisfactory" rating have to be re-evaluated directly in the following year and their contracts are terminated if their rating does not improve. This requirement did not change with the 2011 law. Yet, before 2011, "unsatisfactory" teachers were only dismissed if their rating did not improve in two subsequent evaluations, rather than one.

Teachers in the top two categories also receive access to a rewards and incentive scheme.[17] In contrast, “basic” teachers are barred from applying to this program. Further, teachers in the bottom two categories may be asked to participate in professional development activities before their next evaluation takes place.[18] Yet, crucially, assignment mechanisms for these programs did not change with the 2011 law.

Therefore, the new law affected teachers rated as “basic” (as opposed to “competent” or ”outstanding”) chiefly through the requirement to undergo a renewed evaluation two years later. My analyses are thus able to focus on a comparison of teachers whose performance score suggested a “basic” rating (inducing them to undergo a new evaluation) with a counterfactual situation in which they would have obtained a higher score, before and after the law was passed.

2.4 Identification strategy

This study uses a fuzzy difference-in-difference (“fuzzy DD") estimation strategy. In summary, I exploit that (a) under the new law, teachers below the cutoff were induced to be re-evaluated (in contrast to teachers just above the cutoff), and (b) other effects of a “basic” (in contrast to a higher rating) rating stayed the same over the same period. The estimator moreover accounts for the fact that the law is not adhered to perfectly (in other words, the post-policy jump in the probability of getting evaluated after two years is “fuzzy").

More formally, I estimate a two-stage least squares (2SLS) regression, whose reduced

[17] Chile’s evaluation framework consists of multiple components, which are chiefly the teacher performance evaluation system, Docentemás, the Program for the Variable Individual Performance Allowance (AVDI), the Program for the Accreditation of Pedagogical Excellence Allowance (AEP), and the National System for Performance Evaluation (SNED). AVDI represents a complementary, voluntary, reward system that is open to those municipal instructors rated within the top two of four performance brackets, as determined by Docentemás. AEP, on the other hand, provides an additional, voluntary reward system for all teachers, offering a monetary award to selected candidates, public praise, and the opportunity to apply to the “Maestros” Teacher Network. Lastly, SNED uses national test score data to offer group level incentives to schools (excluding private schools).

[18] These Professional Development Plans (PSPs) are paid for centrally, organized by municipalities, and mainly consist of courses, workshops, and seminars. See Cortés and Lagos (2011) for a detailed description of PSPs and related descriptive statistics, in Spanish. See Lombardi (2019) for an evaluation of their effectiveness.

form is given in Equation 2.1, as follows.[19]

$$Y_{j(t+x)i} = \beta_{RF0} + \beta_{RF1} T_{jt} + \beta_{RF2} TxPost_{jt} + \Gamma_t + \Omega + \boldsymbol{X}_{jti} + \epsilon_{j(t+x)i} \tag{2.1}$$

Reduced-form (RF) Equation 2.1 refers to teacher j, initially evaluated in year t. Here, Y denotes an outcome of interest (e.g., test scores) for teacher j's student i, x years past t. T is an indicator for being below the assignment breakpoint at any point of time, and $TxPost$ is an indicator for being below this breakpoint in the post-policy period (reflecting assignment to re-evaluation as per the continuous evaluation score teacher j received in year t). Γ_t captures year fixed effects; Ω captures commune fixed effects (I omit a commune subscript throughout); $\boldsymbol{X}$ is a vector of teacher and student characteristics measured in baseline year t.[20] The respective first-stage (FS) equation (not shown) is equivalent to Equation 2.1, but now the outcome variable $Y_{j(t+2)}$ reflects a teacher's re-evaluation in year $t+2$, the estimation occurs at the teacher-level, there is hence one observation per teacher in year t, any subscripts i are therefore dropped, and the vector of baseline covariates $\boldsymbol{X}$ excludes student characteristics.

The coefficient of main interest is β_2. In the remainder of the paper, all reported effect sizes (and their standard errors) represent the Wald estimate β_{Wald2}, where $\beta_{Wald2} = \beta_{RF2}/\beta_{FS2}$. Given the (complete lack of) re-evaluations in the pre-period and near-zero re-evaluation rates for the post-period comparison groups (see Section 2.7.1 below), β_{Wald2} is interpreted as a Treatment-on-the-Treated (ToT) effect. In the paper's analysis of heterogeneous effects, I furthermore include interactions between a continuous measure of teacher experience and each of the three variables T, $Post$, and $TxPost$. In the following, β_6 refers to the coefficient on the interaction between $TxPost$ and teacher experience.[21]

I calculate the study's two-stage least-squares Wald estimates through a bootstrap

[19] This notation captures that a fuzzy estimation strategy is equivalent to an instrumental variable (IV) approach. My endogenous variable is teacher (re-)evaluation in year $t+2$, which is instrumented with an indicator for being below the breakpoint, in the post-policy period.

[20] Following common approaches to model student growth trajectories, all student controls also include the quadratic of a child's baseline GPA (cf. Singer and Willett, 2003).

[21] The respective Wald estimate is calculated as follows: $\beta_{Wald6} = (\beta_{RF2} + \beta_{RF6})/(\beta_{FS2} + \beta_{FS6}) - \beta_{Wald2}$.

procedure (with 750 replications) and obtain clustered standard errors by blocking resamples at the teacher-year level. I repeat this procedure for each outcome variable and for the three points in time after a teacher's assignment (that is, $x = 1$, $x = 2$, and $x = 3$).

A final comment is in order as the availability of a continuous assignment variable with a cut-off rule may have—misleadingly—pointed to a simple regression-discontinuity (RD) strategy. However, recall that a regression-discontinuity strategy would not account for the fact that other Chilean programs use the same cut-off to determine eligibility. Additionally, a simple RD approach assumes that teachers' assignment to treatment is as good as random at the threshold score, which implies that teachers are not able to manipulate their scores around this cut-off. Finally, an earlier version of this paper pursued a fuzzy difference-in-discontinuities (or fuzzy difference-in-RD estimator)—it was abandoned, due to lack of statistical power.

2.5 Data

For the years 2005 to 2015, I use teacher-classroom links to match administrative data for the universe of elementary teachers in Chile's public schools, all teacher evaluations conducted in these years, administrative records for the universe of Chilean students, and results on standardized test scores (in mathematics and language) for all Chilean fourth graders attending public schools. More precisely, I combine data from five different sources.[22] The first data source is the "Ideoneidad Docente" data-base, which is maintained by Chile's Ministry of Education. The data-base includes detailed, administrative information on the population of Chilean teachers such as information on a teacher's age and gender, a teacher's years of experience in the school system, contractual details (such as the number of working hours), information on a teacher's training (such as subject specialization and the training institution), identifiers for the school and grade level a teacher taught in a given

[22] If not indicated otherwise, data-sources are in the public domain and can be downloaded from a website maintained by the Education Ministry's "Centro de Estudios" (2016). Data-sets are merged by using unique school identifiers, information on grade levels and classes, and unique (codified) teacher identifiers.

year, and information on the school (such as whether a school is located in an urban or in a rural area).

Second, the above data is merged with a data-set containing information on whether a teacher participated in Docentemás in any given year between 2005 and 2015. This data-set also includes detailed information on each teacher's final performance rating, the continuous performance score, and her rating on each of the four evaluation components. The study's third data-source consists of the "Asignatura por Docente" data-set for 2005-2015, which provides information on the class(es) and subject(s) a teacher taught in a given year.[23] Fourth, the study combines administrative information on the universe of Chilean students, their absentee rate (as a percentage of school days), whether they repeated a given school year, and their end-of-year grade point average (GPA).

Fifth, student learning outcomes are measured using Simce exam scores.[24] The Sistema de Medición de la Calidad de la Educación (Education Quality Measurement System, in short: "Simce") was first introduced in 1988 and represents a mandatory, full-cohort, standardized exam administered at the end of the school year (across private, subsidized, and public schools). As of 2015, Simce has covered a wide array of subjects and levels, but most notably fourth-grade mathematics and (Spanish) language in every consecutive year since 2005. Simce data-sets include information on the student's gender, school, and class, and additional student demographics (such as the mother's highest level of education and the family's level of income). Given the salience of Simce scores in Chile, I do not transform them to standard deviations. However, results remain easily interpretable as Simce test-scores are scaled to a standard deviation of 50.

For each year, Simce assessments are also complemented by a student, a parent, and a teacher survey. For a subset of years, these surveys provide comparable measures of potential mediating factors: teaching effort, teachers' level of caring, and teacher beliefs. For

[23] This data-set is not in the public domain. The Ministry asks researchers to undergo a standardized process to receive access to the data-set.

[24] The student-level version of this data-set is not in the public domain. The Ministry asks researchers to undergo a standardized process to receive access to the data-set.

the years 2012 through 2015, I construct an index of student-reported teaching practices, or “effort”. More specifically, I calculate an average over six survey questions that seek to capture a teacher’s classroom behaviors.[25] Moreover, for 2011 to 2014, I use a measure that asks parents to rate the extent to which their child’s head teacher cares about her students.[26] As there is no head teacher identifier, in my analyses of mechanisms, I assume that in fourth grade, a teacher is considered the head teacher if she teaches both math and language.[27] Finally, for each year from 2005 to 2014, surveys ask teachers to state their beliefs concerning students’ future educational attainment.[28]

2.6 Sample characteristics and internal validity

2.6.1 Sample

Table 2.1 provides summary statistics for the study’s sample of teachers at “baseline" (that is, the year of their initial evaluation, t), and the analysis sample of teachers observed teaching grade four students two years later (in year $t+2$). This includes teachers who were initially evaluated between 2005 and 2013, and potentially re-evaluated between 2007 and 2015. Appendix Table B1 presents the respective descriptives for years $t+1$ and $t+3$.[29]

[25] Students answer in four categories: “Fully agree”, “agree”, “disagree", “very much disagree”. Students are asked about whether their teacher 1) reviews exercises, 2) reviews homework, 3) explains something repeatedly if someone asks for it, 4) continues to explain until everyone understands, 5) explains in class how tests were marked, 6) corrects the school book’s exercises in class. Results (available upon request) are robust to using an alternative index from a principal component analysis instead (with a polychoric correlation matrix, extracting the first joint component from the six items).

[26] Rated from 1 or “very unsatisfied” to 7 “very satisfied”.

[27] Approximately, 90 percent of the study’s sample of fourth graders have the same math and language teacher. I drop the remaining observations when analyzing mechanisms. In 2015, students were asked separately about their math and language teacher’s classroom behavior. For this year, I average the student responses across subjects.

[28] Teachers choose one of the following six categories: 1) Will not complete eighth grade, 2) will complete eighth grade on the technical-professional track, 3) will complete eighth grade on the humanist-scientific track, 4) will complete a technical degree, 5) will complete a university degree, 6) will complete postgraduate studies.

[29] In 2016, Chile introduced major changes to teachers’ career pathway, including in the way teacher evaluations are used (*Ley 20.903*). This suggests that data for 2016 and thereafter should not be used for the present analysis. At the same time, I do not have access to data for these years. Teachers who were initially

I restrict all analyses to teachers who were initially evaluated in elementary and I drop those teachers who would have been too old to be eligible for re-evaluation two years later.[30] I also drop the small share of approximately 1.8 percent of teachers with a performance score that would have suggested an “unsatisfactory" rating.[31] This approach renders 31,400 teacher-by-year observations (22,435 for the pre-policy period and 8,965 for the post-policy period). Of those, 7,458 represent teachers of grade four students in language or mathematics, two years after the teacher’s assignment to re-evaluation.[32] This mapping of teachers to their students results in 157,153 year-teacher-student observations.[33]

2.6.2 Threats to internal validity

In the following, I investigate—and present evidence against—five potential threats to the study’s internal validity: differential attrition based on a teacher’s assignment status (as per her initial evaluation score in year *t*); imbalance of observable teacher characteristics across teachers “assigned to treatment” and their comparison group; potential sorting of students to (or away from) teachers who are assigned to be re-evaluated; whether, with the policy change, teachers manipulated their assignment status more (/less); and whether the two groups of teachers (those to be “assigned to treatment” and their comparison group) exhibited differential pre-trends.

evaluated in 2013 (and their students) are thus not observed in $t + 3$.

[30] Teachers within three years of the retirement age are not required to be evaluated.

[31] It is unclear whether, due to the new law, informal re-evaluation criteria may have changed for these teachers. I do not use other criteria to drop teachers (such as the teacher’s workload), as these may have changed post-assignment.

[32] These 31,400 teacher-by-year observations comprise 22,066 unique teachers, of whom 6,844 teach fourth-grade language or mathematics.

[33] These year-teacher-student observations comprise 119,382 unique students. Students may be observed twice, either if math and language are taught by different teachers, in a given year, or if students repeat fourth grade.

Table 2.1: *Sample characteristics and validity checks*

	2005-2010		2011-13		
	Below	Above	Below	Above	DD
Attrition					
In sample in t+1	0.18	0.21	0.15	0.21	-0.02 (0.01)*
In sample in t+2	0.17	0.21	0.15	0.21	-0.02 (0.01)
In sample in t+3	0.16	0.2	0.11	0.1	0.01 (0.01)
n	8,192	14,243	1,599	7,366	31,400
Teacher Baseline Characteristics					
Gender: Female	0.75	0.83	0.77	0.86	-0.03 (0.03)
Contract hours	38.56	38.28	38.11	37.35	0.67 (0.50)
Works in yet another school	0.07	0.06	0.04	0.02	0.01 (0.02)
Years in service	19.76	18.66	14.23	14.96	-0.83 (0.79)
n	1,886	3,612	279	1,681	7,458
School's baseline reading score†	241.68	247.83	249.85	254.95	-0.35 (1.53)
School's baseline math score†	230.4	236.01	238.96	246.52	-1.74 (1.72)
n	1,387	2,928	235	1,533	6,083
Student Baseline Characteristics					
GPA	5.89	5.95	5.87	5.91	0.03 (0.02)
Repeated in baseline year	0.03	0.03	0.04	0.04	0.00 (0.00)
Attendance	92.64	92.97	90.78	91.65	-0.24 (0.25)
Gender: Female	0.49	0.49	0.48	0.49	-0.00 (0.01)
n	36,410	79,242	5,391	36,110	157,153
Household income (pesos)††	265977	278862	343669	328037	-1242.60 (12244.54)
Mother's edu. (years)††	9.85	10.18	10.76	10.81	-0.08 (0.11)
n	29,637	66,198	4,655	31,666	132,156

Notes. "Teachers" include all unique year-teacher observations and may thus repeatedly include individual teachers over time. "Students" include all unique year-teacherstudent observations and may thus include up to two observations per student and year (if math and reading are taught by different teachers, in a given year). Teacher and student baseline characteristics refer to the analysis sample observed at $t+2$. Appendix Table B1 reports on the sample observed for $t+1$ and $t+3$. "Below" and "Above" refer to teachers below or above the cut-off, respectively. t refers to the year of the initial evaluation. All variables measured in t, if not denoted otherwise. † denotes variables available for fewer observations (and not included as covariates). †† denotes variables measured at follow-up (and not included as covariates). Note that the 2013 sample is not followed up in $t+3$. "DD" refers to a difference-in-difference estimate as described in Section 2.4 (excluding control variables but including commune-level fixed effects). Standard errors in parentheses. For student-level characteristics, standard errors are clustered at the year-teacher level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Differential attrition

Table 2.1's last column follows the paper's difference-in-difference strategy (as described in Section 2.4 above, with the exclusion of covariates). The Table's top panel reports whether the introduction of the law coincided with systematic changes in "attrition" rates—i.e., whether students became systematically less (/more) likely to be taught by a "basic" mathematics or language teacher. Table 2.1's findings suggest that, in the post-policy period, "basic" teachers became slightly less likely to teach a fourth-grade class in mathematics or language in the year after evaluation scores are released (by two percentage points, statistically significant at the 0.1 level). If these weaker teachers are thus systematically removed, Equation 2.1 may be slightly under-estimating the true effect of teachers' re-evaluations, for year $t+1$. For the year of the re-evaluation ($t+2$) and the year thereafter ($t+3$), in contrast, I do not find evidence of systematic removal (or assignment) of "basic" teachers to fourth-grade math and language classrooms, as the new law was introduced.

Baseline balance for teachers

For fourth-grade Simce teachers, there are only negligible differences in teachers' gender or age, their contract hours, their years of experience, and in the percentage of teachers who work in more than one school. At baseline, I also find no differences in teachers' average school-level Simce scores (whether in fourth-grade math or language). As an exception, three years post assignment, assigned teachers were slightly less experienced, by 2.4 years (significant at the 0.05 level; see Appendix Table B1). This difference is not confirmed for the remaining two samples. This finding therefore appears to reflect multiple hypothesis testing, rather than systematic differences. Yet, I also control for these teacher characteristics, including a teacher's years of experience, in the vector of baseline covariates.

Baseline balance for students

To investigate the potential of systematic sorting of students, Table 2.1 also includes descriptive statistics for teachers' fourth-grade Simce-taking students, two years post-assignment (yet measured at baseline, in year *t*).[34] The table also presents additional information on student demographics (household income and the mother's highest level of education), even though this information is not available for all students, it is measured at the time of follow-up, and in all of the paper's regression analyses, these variables are therefore not included as covariates.

I find no support for the hypothesis that schools engage in sorting of students to (or away from) teachers who are assigned to be evaluated. None of the tests point to differences in student characteristics (at the 0.1 level). Moreover, point estimates are close to zero, with tight error bands, suggesting that students' prior academic achievement, grade retention, attendance, gender, household income, and maternal level of education are balanced as the difference among groups below and above the assignment threshold is compared across the pre- and post-policy periods.

Differential manipulation of assignment status

Another concern revolves around whether, with the policy change, teachers close to the threshold for a "basic" rating were systematically assigned to more (/less) lenient performance ratings. As teachers, their peers, and principals determine 40 percent of the performance score (through self- and peer-evaluations, as well as reference-reports), there may have been an increase (/decrease) in the share of teachers whose score—and thus treatment assignment—was manipulated. Two facts alleviate this concern.

First, recall that the remaining 60 percent of a teachers' continuous score is based on her portfolio, which is rated centrally, anonymously, and *after* the remaining three components have been scored. For a teacher, her peers or the principal, it is thus impossible to know whether the composite score will be close to the cut-off.[35]

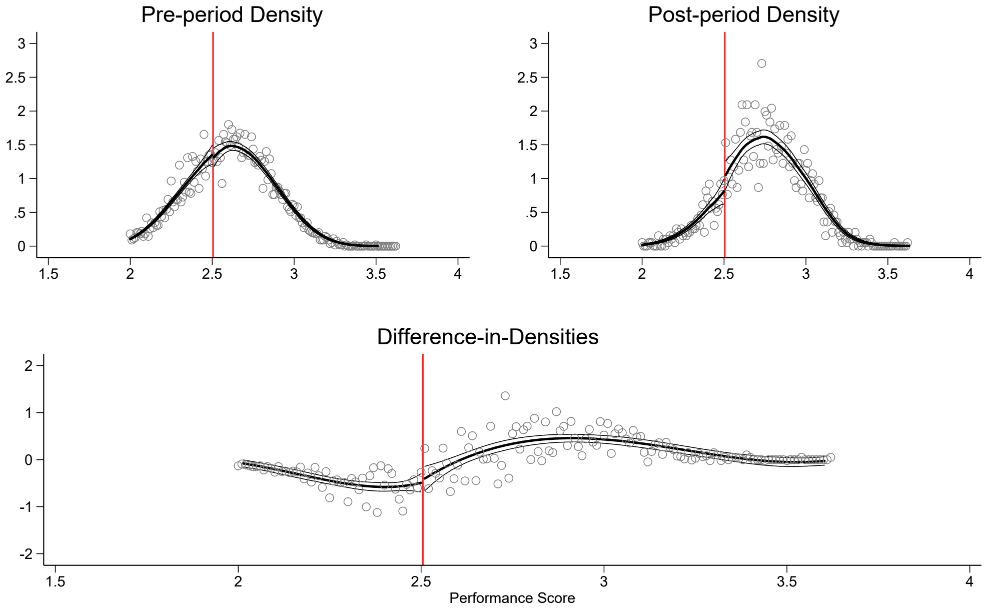
Secondly, in Figure 2.1, I build on work by McCrary (2008) to assess empirically whether

[34] Appendix Table B1 provides the respective information for the samples one year and three years past baseline.

[35] See Lombardi (2019), for a similar argument.

score manipulation around the cut-off score differed across the pre- and post-policy periods.

Figure 2.1: *McCrary plots*



Note: Vertical lines indicate the recommended cutoff score.

The figure's top-panel shows McCrary plots for the pre-period (left) and the post-period (right). These plots are generated by calculating a finely-gridded histogram, which is then smoothed using local linear regression, separately on either side of the breakpoint. A formal test (McCrary, 2008) on the difference-in-densities around the breakpoint does not reject the null of equal densities on both sides, for either period (at the 0.01 level). Further, this paper's identification strategy simply posits that, if present at all, the extent of manipulation remained unaffected by the policy change. In the bottom panel, I extend McCrary's (*ibid.*) method by calculating the difference-in-difference of densities, for common bin-sizes of 0.01 points[36], and smoothing over the histogram thereafter (separately, for both sides of the breakpoint).[37] The bottom panel illustrates how the difference in densities around the

[36] Teacher evaluation scores are reported in increments of 0.01 points.

[37] To my best knowledge, this is the first study presenting a McCrary plot for the difference-in-differences of densities. However, I do not calculate optimal bin sizes and deviate from McCrary's (2008) method of choosing the optimal bandwidth. I choose a bandwidth of 0.2 points and a bin size of 0.01 points, as in the remainder of

breakpoint remains constant over time (not significantly different from zero at the 0.1 level). In summary, this graph (and the respective test of a difference-in-difference of densities) thus shows that a difference-in-differences approach alleviates concerns regarding manipulation around the cut-off.

Common trends

As with any difference-in-difference estimation, the identifying assumption is that the average change in the comparison group's outcomes represents the counterfactual change in the treatment group's outcomes (in the absence of treatment). While not directly testable, I present pre-policy trends in outcome variables, for teachers assigned to a basic rating vs. teachers assigned to a higher rating (in year $t + 2$). Appendix Figure B1 shows that, in the pre-policy period, the explained (left panel) and unexplained (right panel) portions of test score variance follow parallel trends, across these two groups of teachers. I therefore conclude that there is no evidence to suggest a violation of the common trends assumption.

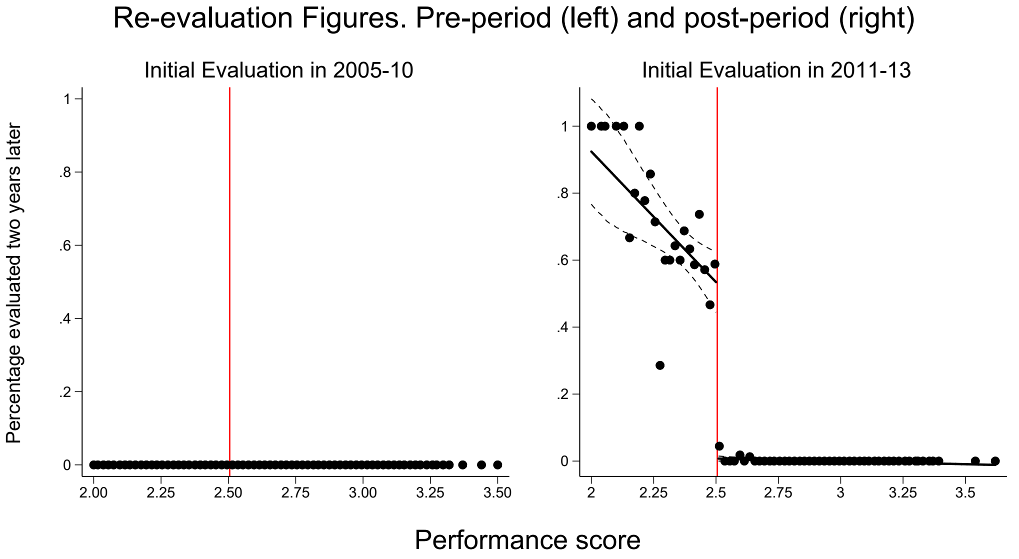
2.7 Results

2.7.1 First stage results

Figure 2.2 provides evidence for the validity of the study's first stage. Each point represents the share of teachers being re-evaluated after two years, in score bins with a width of 0.02 score points. The solid line plots predicted values, with separate linear trends estimated on either side of the basic vs. competent (or better) cut-off. This threshold is indicated by the red, vertical line. The dashed lines show 95 percent confidence intervals. In the pre-period (left panel), the percentage of teachers who are newly evaluated in year $t + 2$ is consistently zero. Thus, by including the pre-period, the proposed estimator solely differences out potential effects that occurred around the same threshold (for example, through eligibility

the paper. For consistency with McCrary's (*ibid.*) method, I include a fourth-order polynomial on both sides of the breakpoint. I thank Ugo Troiano for helpful comments.

Figure 2.2: *First stage*



Notes: Each point represents the share of teachers being re-evaluated in score bins of width 0.02 score points. The solid line plots predicted values, with separate linear trends estimated on either side of the basic vs. competent threshold. This threshold is indicated by the vertical line. The dashed lines show 95 percent confidence intervals. No predicted values or confidence intervals shown for the pre-period as the share is consistently zero.

for incentives or training, rather than an effect of re-evaluations).

In the post-period (right panel), as expected, a large jump in the probability of re-evaluation occurs around the breakpoint. Teachers' predicted share of re-evaluation just to the left of the threshold (suggesting a "basic" rating) is 63 percent; in contrast, the percentage remains close to zero once the threshold is crossed (0.2 percent). Note that there is great variance with respect to compliance (or the level of "fuzziness") among the assigned teachers. Yet, in addition to this visual evidence, the formal estimate of Equation 2.1 also confirms the strength of the first stage relationship—the F statistic for a test of $\beta_{FS2} = 0$ is 3843.4, for the sample of teachers teaching fourth-grade two years after the assignment. This formal estimate suggests that the law increased the probability of re-evaluation by 62.8 percentage points.[38]

[38] Results for those teachers observed teaching fourth-grade in $t + 1$ and $t + 3$ are similar and available upon request.

2.7.2 Effects on student learning

Table 2.2 shows the study's main results, for the investigation of effects on student learning. In Table 2.2, coefficients for β_1, year and commune fixed effects, and the vector of teacher and student characteristics are omitted. Models in odd-numbered columns do not account for potentially heterogeneous effects by teacher's level of work-experience. Even-numbered columns refer to models that interact the treatment with a teacher's years of work-experience. Recall that β_{Wald2} captures the main ToT effect of a teacher's re-evaluation, whereas β_{Wald6} reflects the additional ToT effect, times the teacher's years of experience in year t.[39] Recall also that, in each year and subject, Simce scores are scaled to a standard deviation of 50.

Table 2.2: *ToT effects on student learning*

	t+1		t+2		t+3	
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
β_{Wald2}	-3.045 (4.189)	-6.687 (5.668)	-2.040 (2.991)	-5.167 (4.730)	0.484 (3.887)	-1.274 (6.173)
β_{Wald6}		0.117 (0.264)		0.210 (0.256)		0.032 (0.301)
n (teachers)	6991	6911	6643	6565	5417	5336
n (students)	142515	142515	133013	133013	110000	110000
Language						
β_{Wald2}	-4.328 (3.591)	-8.522 (4.823)*	-1.966 (2.727)	-6.194 (4.224)	2.287 (3.486)	-4.232 (5.311)
β_{Wald6}		0.133 (0.224)		0.337 (0.260)		0.418 (0.274)
n (teachers)	7158	7076	6869	6785	5645	5556
n (students)	144868	144868	136938	136938	112986	112986

Notes. In odd columns, β_{Wald2} captures the ToT effect of a teacher's re-evaluation. In even columns, β_{Wald6} captures the interaction (ToT) effect between a teacher's re-evaluation and her work experience. Not reported: Main effect, year fixed effects and a vector of baseline teacher and student characteristics (teacher's gender, age, contract hours, employment in another school, years of service, baseline school-level average Simce scores in math and reading; students' GPA and its square, attendance, retention). Bootstrapped standard errors in parentheses (750 draws), clustered at the teacher-year level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

As shown in Column (5), I find that students who were taught by a teacher who was re-evaluated one year prior did not perform differently, compared to their peers, whether in math or reading. Column (3) also does not lend support for the hypothesis that there is a detrimental effect of teacher evaluations on student test scores. Column (1) reports findings for the year prior to a teacher's evaluation, one year after the "treatment" was assigned ($t + 1$). In this year, teachers may have changed their behavior once they learned about their

[39] Given the two-stage least-squares set-up, I do not report R^2.

treatment status (Ashenfelter, 1978). Yet, for both subjects, I do not find such an effect.

Columns (2), (4), and (6) assess whether these results differ for teachers with fewer (or more) years of work experience. The results do not support such a phenomenon; the coefficients of β_{Wald6} are statistically indistinguishable from zero. Yet, for language and year $t+1$, I find a negative effect among new teachers (approximately 0.17 standard deviations in expectation, significant at the 0.1 level). This effect is of similar size for mathematics but not statistically significant. In summary, I therefore conclude that teacher's re-evaluations did not lead to increased teacher productivity (as measured by student test scores). I find that this observation is independent from a teacher's level of work experience.

2.7.3 Effects on teachers and teaching

Table 2.3 reports on effects on teachers' (student-reported) teaching behaviors, (parent-reported) levels of caring, and (self-reported) beliefs in their students' future educational attainment. All measures are standardized. All models follow Equation 2.1; they include a vector of baseline teacher and student characteristics, commune and year fixed effects, and cluster standard errors at the teacher-year level.

Table 2.3: *ToT effects on teaching behaviors, caring, teacher beliefs*

	t+1		t+2		t+3	
	(1)	(2)	(3)	(4)	(5)	(6)
Practices						
β_{Wald2}	-0.110 (0.110)	-0.113 (0.164)	-0.220 (0.067)***	-0.261 (0.117)**	-0.295 (0.088)***	-0.215 (0.130)*
β_{Wald6}		0.000 (0.007)		0.004 (0.006)		-0.007 (0.008)
n (teachers)	2291	2267	3183	3142	2511	2477
n (students)	45207	45207	61689	61689	50315	50315
Caring						
β_{Wald2}	-0.242 (0.112)**	-0.276 (0.169)	0.050 (0.091)	0.059 (0.138)	0.006 (0.086)	0.072 (0.160)
β_{Wald6}		0.001 (0.008)		-0.000 (0.006)		-0.004 (0.009)
n (teachers)	2065	2044	2143	2117	1912	1889
n (parents)	40835	40835	41725	41725	39154	39154
Beliefs						
β_{Wald2}	0.106 (0.177)	-0.165 (0.247)	0.109 (0.164)	0.138 (0.234)	-0.085 (0.286)	0.076 (0.478)
β_{Wald6}		0.016 (0.011)		-0.002 (0.010)		-0.010 (0.029)
n (teachers)	5853	5788	4801	4749	4028	3972

Notes. In odd columns, β_{Wald2} captures the ToT effect of a teacher's re-evaluation. In even columns, β_{Wald6} captures the interaction (ToT) effect between a teacher's re-evaluation and her work experience. Not reported: Main effect, year fixed effects and a vector of baseline teacher and student characteristics (teacher's gender, age, contract hours, employment in another school, years of service, baseline school-level average Simce scores in math and reading; students' GPA and its square, attendance, retention). Bootstrapped standard errors in parentheses (750 draws), clustered at the teacher-year level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.3 suggests negative effects on teaching practices in the year of a teacher's re-evaluation, and in the year thereafter (of 0.22 and 0.30 standard deviations, respectively; significant at the 0.01 level). The table also documents negative effects on teachers' levels of caring, in the year after the teacher's assignment (of 0.24 standard deviations; significant at the 0.05 level). I do not find other effects for the remaining year-outcome combinations. Moreover, none of these results differ by teachers' level of work experience.

2.7.4 Robustness of findings

I present three types of robustness checks. I present evidence for the absence of effects in the pre-period (i.e., $t-1$, the year prior to a teacher's potential assignment). I moreover re-estimate the study's main results by adding a group-specific, linear time trend for "basic" teachers to Equation 2.1. Finally, I re-estimate the study's main results by slightly modifying the assignment indicator. Results from these checks are presented in Table 2.4 below.

Table 2.4: *Robustness Checks*

	Falsification	Group-specific time trends			Re-defining assignment		
	t-1	t+1	t+2	t+3	t+1	t+2	t+3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math							
β_{Wald2}	-0.982 (4.182)	-8.698 (5.871)	2.593 (4.648)	-2.567 (4.932)	-4.890 (3.342)	-1.480 (2.724)	-2.090 (3.305)
n (teachers)	4313	6083	5726	4627	6083	5726	4627
n (students)	100049	142515	133013	110000	142515	133013	110000
Language							
β_{Wald2}	1.748 (3.510)	-13.868 (5.243)***	3.027 (4.341)	-0.009 (4.419)	-5.784 (3.119)*	-0.752 (2.539)	-0.524 (2.859)
n (teachers)	4410	6218	5857	4853	6218	5857	4853
n (students)	101649	144868	136938	112986	144868	136938	112986
Practices							
β_{Wald2}	-0.016 (0.109)				0.027 (0.096)	-0.161 (0.065)**	-0.268 (0.069)***
n (teachers)	1027				2296	2985	2385
n (students)	19001				42016	55654	44025
Caring							
β_{Wald2}	-0.147 (0.127)				-0.157 (0.095)*	0.021 (0.082)	0.015 (0.075)
n (teachers)	955				2044	2031	1814
n (parents)	16946				37566	37840	34188
Beliefs							
β_{Wald2}	0.011 (0.161)	0.123 (0.197)	-0.051 (0.186)	-0.115 (0.223)	-0.012 (0.120)	-0.024 (0.132)	0.011 (0.180)
n (teachers)	4653	6536	5422	4603	6536	5422	4603

Notes: β_{Wald2} captures the ToT effect of a teacher's re-evaluation.
Column (1) reports on effects the year prior to assignment. Columns (2) to (4) add a time-trend for teachers with a "basic" evaluation score to Equation 2.1.
Group-specific time trends omitted for outcomes with data for only three years of pre-policy data, or less.
Columns (5) to (7) base a teacher's assignment on her final evaluation rating, not on the underlying evaluation score.
Not reported: Main effect, year fixed effects and a vector of baseline teacher and student characteristics (teacher's gender, age, contract hours, employment in another school, years of service, baseline school-level average Simce scores in math and reading; students' GPA and its square, attendance, retention).
Bootstrapped standard errors in parentheses (750 draws), clustered at the teacher-year level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Falsification test

Column (1) of Table 2.4 presents results from a falsification test. In this test, I estimate Equation 2.1 for students one year *before* their teacher was subjected to her initial evaluation (and thus potentially assigned to be re-evaluated). I do not find evidence for "effects" of a teacher's later assignment. This result further corroborates the findings from Section 2.6.2, above, and alleviates concerns of differential pre-trends.

Robustness under inclusion of time trends

Columns (2) to (4) of Table 2.4 present a re-estimation of Equation 2.1 with the inclusion of a group-specific linear trend, for teachers with an evaluation score that suggests a "basic" rating. This check is available for learning outcomes and teachers' beliefs only; the remaining two outcomes only provide data for two (/three) pre-policy years, respectively. More specifically, I maintain Equation 2.1's year fixed effects and add an interaction of T with a continuous year indicator.

The results from this analysis confirm the findings reported in Section 2.7.2 (see Table 2.2, odd columns) for math, and for language in the year of and the year after a teacher's re-evaluation. In the year prior to the re-evaluation, however, we now observe negative effects on language (of .27 standard deviations). All remaining effects on instruction, teachers' level of caring, and beliefs are indistinguishable from zero.

Robustness to alternative treatment assignment

So far, the article's analyses have used the continuous score to reconstruct whether a teacher was to be assigned for re-evaluation. Columns (5) to (7) of Table 2.4 present findings from a re-estimation of Equation 2.1 if a teacher's actual performance rating is used for this purpose instead.[40] These results document a negative effect (of 0.11 standard deviations) on language learning in year $t + 1$, the year after assignment and prior to re-evaluation.

[40] Recall that a local commission may object to a teacher's score and override her rating. Note how this alternative assignment definition may thus not be entirely exogenous.

The respective effect for math is of similar size but statistically insignificant. For that year, they also suggest negative effects (of 0.16 standard deviations) on teachers' levels of caring. Finally, the results point to negative impacts on teaching practices both in the year of and in the year after a teacher's re-evaluation (of 0.16 and 0.27 standard deviations, respectively). All remaining coefficients are indistinguishable from zero.

2.7.5 Bounding of positive ITT effects on student learning

Are the above null-findings on student learning precise enough to rule out positive effects of Chile's policy, which mandates repeat teacher evaluations? To investigate this question, I now switch the study's focus to intent-to-treat (ITT) estimates. I repeat the above bootstrap procedure, plot the distribution of ITT coefficients (from 750 draws), and report on the 95th percentile. I interpret this value as the upper bound of ITT effects, and rule out higher impacts. To gain further precision, in doing so, I follow de Ree *et al.* (2018) and pool observations across subjects.

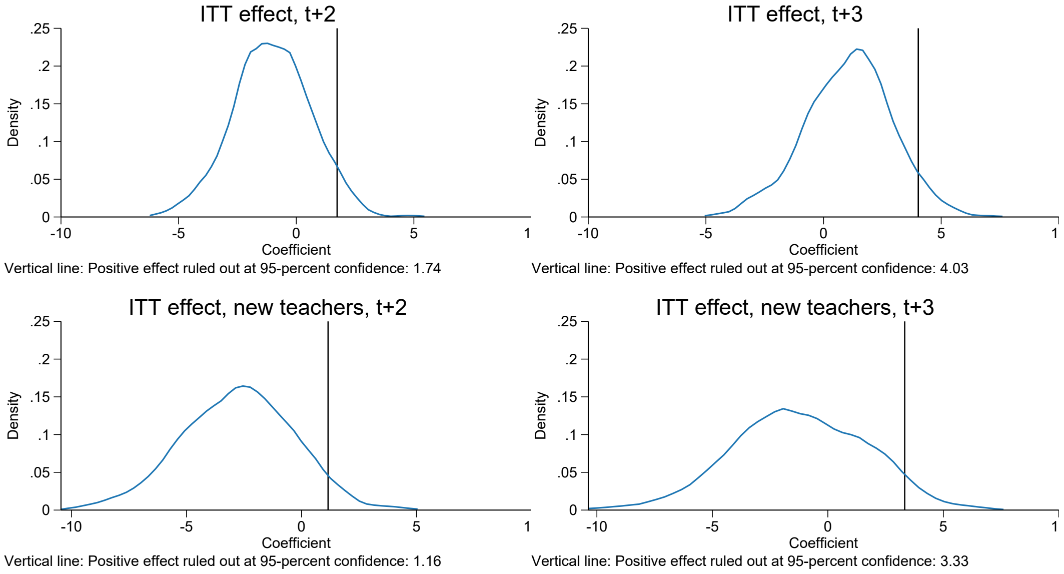
Figure 2.3 reports on the results from this bounding exercise. The top panels reports on the ITT effect in the year of the evaluation (left panel) and in the year after the evaluation (right panel). The bottom panels allow for heterogeneity in impacts and report on the corresponding ITT effects for teachers with just one year of work experience.

Positive ITT effects are ruled out precisely. With 95 percent confidence, the results reject effects larger than 0.03 standard deviations for the year of the evaluation, and of 0.08 standard deviations for the year after the evaluations. For teachers who just started teaching, and may be expected to be most receptive of personal development, these bounds are even smaller (0.02 standard deviations and 0.06 standard deviations, respectively).

2.8 Conclusion

This study offers quasi-experimental evidence on the effects of formative performance evaluations on teacher effectiveness and child learning. In summary, I cannot conclude that Chile's repeat performance evaluations lead to substantial gains in student achievement,

Figure 2.3: *Bounding of positive ITT effects on student learning*


ITT effect, t+2

Vertical line: Positive effect ruled out at 95-percent confidence: 1.74

ITT effect, t+3

Vertical line: Positive effect ruled out at 95-percent confidence: 4.03

ITT effect, new teachers, t+2

Vertical line: Positive effect ruled out at 95-percent confidence: 1.16

ITT effect, new teachers, t+3

Vertical line: Positive effect ruled out at 95-percent confidence: 3.33

Notes: This figure provides kernel density plots of ITT coefficients, from a bootstrap with 750 draws, clustered at the teacher-year level. Vertical lines indicate the 95th percentile. 'New teachers' refers to teachers with one year of experience. Regressions include year fixed effects and a vector of baseline teacher and student characteristics (teacher's gender, age, contract hours, employment in another school, years of service, baseline school-level average Simce scores in math and reading; students' GPA and its square, attendance, retention).

one year after a teacher is assigned to be evaluated. For the year of the evaluation, I also do not find effects on student learning. Positive impacts are ruled out precisely. Instead, the study results suggest that concerns about detrimental effects may be at least partly warranted. Some specifications point to negative effects on language learning, in the year after a teacher's assignment. The paper moreover documents decreases in teachers' level of caring, for the same year. I also observe additional negative effects on teaching practices in the year of and in the year after a teacher's evaluation. I do not detect impacts on teachers' beliefs in their students' future educational attainment.

This study isolates the effect of *repeat* evaluations; it cannot speak to the effects of a teacher's initial evaluation. Yet, for policy makers considering the introduction of a national system, this is arguably the more important question to consider: Will teachers' performance and student learning improve as teachers are regularly subjected to evaluations?[41] Taken together, in evaluating the impact of repeat evaluations, this study aims to provide first

[41] Compare to Taylor and Tyler (2012, 3629), who also stress this point.

evidence on this question—for a comprehensive, standards-based teacher evaluation system that has been described as a role model for other countries (Bruns and Luque, 2014). To my best knowledge, it is only the second rigorous study on the effects of formative performance evaluations, and the first analysis under a well-established evaluation system that operates at national scale.

As discussed by Taut *et al.* (2011), Chilean policy makers regularly re-consider whether "Docentemás" is worth its cost and whether the system should be expanded to private schools (Educación 2020, 2013). Moreover, given the scale and nature of the investigated program, even decision makers in other public sectors may look to the example of Docentemás as staff performance evaluation systems are (re-)considered. In the light of these debates, this article casts doubt on the use of repeated formative evaluations as a means to improve employee productivity.

Chapter 3

Do Students Benefit from Blended Instruction? Experimental Evidence from Public Schools in India[1]

3.1 Introduction

High-quality teaching is a key determinant of student success. A growing body of literature documents how a large proportion of classroom-to-classroom variance in student performance can be attributed to teachers' instructional practices (Araujo *et al.*, 2016; Azam and Kingdon, 2015; Bau and Das, 2017; Buhl-Wiggers *et al.*, 2017).[2] Beyond test scores, teaching quality is also a main driver for the development of socio-emotional skill (Jackson, 2018) and other long-term life outcomes (Rivkin *et al.*, 2005; Chetty *et al.*, 2014; Jackson *et al.*, 2014). At the same time, teachers in many less-developed countries may lack the necessary skills to teach effectively and use teaching methods ill-matched to their students' diverse needs (Bietenbeck *et al.*, 2018; Bold *et al.*, 2017)—even in countries with comparatively high teacher

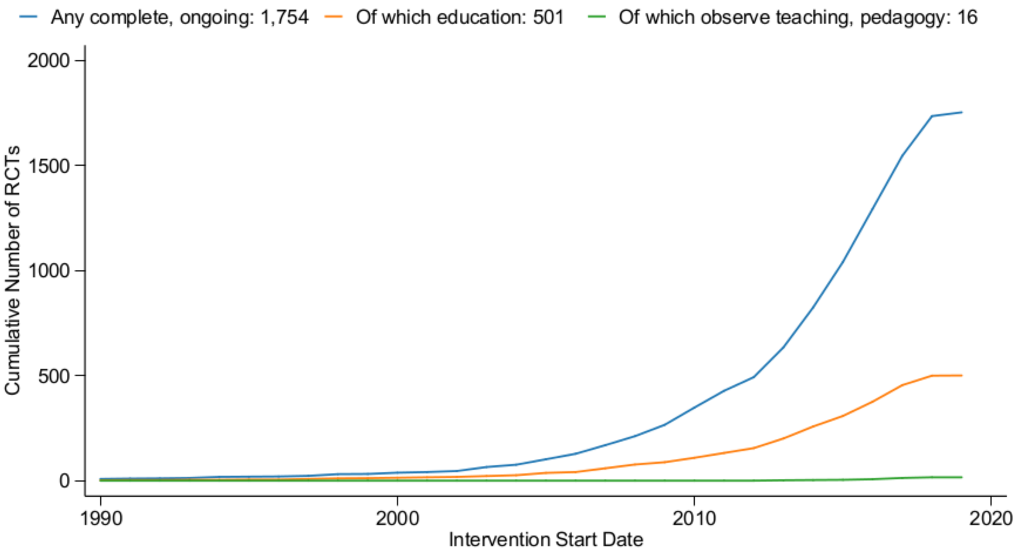
[1] Single-authored.

[2] In contrast, observable characteristics of teachers (rather than their teaching practices), are often considered a poor predictor of student learning (*ibid.*).

pay (de Ree *et al.*, 2018; Ramachandran *et al.*, 2018).

Figure 3.1 documents how, despite this importance of instructional quality (and the potential lack thereof), prior experimental studies on the economics of education have largely ignored teaching practices and pedagogy. Amidst a "Global Learning Crisis" in less-developed countries—where high enrollment numbers have not coincided with increases in student skill (The World Bank, 2017a)—there has been a large increase in the number of rigorous research on "what works" to improve student learning. Out of the 1,754 complete and ongoing trials registered at the AEA registry, 501 study education (29 percent). Yet, of those, only 16 measure outcomes relating to pedagogy or teaching practices. It is within this context that I set out to conduct a large, cluster-randomized trial, to study the effects of an intervention that aims to generate improvements in instructional quality.

Figure 3.1: *Education RCTs on the AEA Trial Registry ignore teaching, pedagogy*



Notes. This figure reports on the number of complete and ongoing randomized experiments registered at socialscienceregistry.org (as of 18 April 2019).

In this article, I present experimental evidence on the effects of a computer-assisted educational program that encourages teachers to blend their instruction with high-quality video materials. It does so by providing 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, 2017). 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 0.14 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 0.08 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 suggestive evidence that the negative effects in mathematics are driven by grade-9 students. For grade 9, in mathematics, students in ICT schools performed 0.18 standard deviations worse than students in Control schools. The difference to students in Workbook schools is 0.14 standard deviations. A non-parametric investigation of heterogeneous effects moreover shows that the impacts hold for a wide range of baseline performance levels. Finally, I find large 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 differed by 0.57 standard deviations in math, and 0.25 standard deviations in 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.[3] 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

[3] This differs from other technology interventions that substitute in-school teaching with one-on-one software (e.g., Araya *et al.*, 2019; Banerjee *et al.*, 2007; Carrillo *et al.*, 2010b; Lai *et al.*, 2015; Linden, 2008; Muralidharan *et al.*, 2019; Taylor, 2018) or audio and video materials (e.g., Fabregas, 2018; Jamison *et al.*, 1981; Johnston and Ksoll, 2017; Naik *et al.*, 2020; Navarro-Sola, 2019; Seo, 2017). It also differs from technology interventions that provide additional instruction outside of school, including through phone- or tablet-based applications, or through distance instruction. For recent overviews on the effectiveness of educational technology, including these approaches, see Bulman and Fairlie (2016) and Escueta *et al.* (2017).

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, 2020; 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.*, 2018a; 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.*, 2019; 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 (for an overview, see Escueta *et al.* (2017)). 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 3.2 describes the study's context and provides intervention details. Section 3.3 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 3.4 provides results and Section 3.5 concludes.

3.2 The program

3.2.1 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).[4] s 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).

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

The study is being performed in partnership with Haryana's State Government and its "Government Senior Secondary Schools" (GSSS).[5] These schools are predominantly rural (80 percent of schools), and teach an exclusively Hindi curriculum. In Section 3.3.2, I provide additional information on observable school characteristics and student performance—for the study sample, for Haryana, and India.

3.2.2 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.[6] They are embedded in a tablet-based application, which organizes the materials along with the textbook's chapters and sub-chapters.[7] 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

[5] 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.

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

[7] 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.

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 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.[8] 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 C1 summarizes the program components and provides additional details on the distribution of responsibilities across Avanti Fellows and the state government.

3.3 Evaluation design

3.3.1 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

[8] 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.

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.

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.[9] 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.

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).[10] Trained observers

[9] 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.

[10] 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

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.3, 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.

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).

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.

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.

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 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.[11]

I scaled the results using a two-parameter Item Response Theory (2PL IRT) model, separately for mathematics and science.[12] 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).[13] In Appendix C.4, I provide additional information on these student assessments, their properties, and on my psychometric approach.

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

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

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

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

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.

3.3.2 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 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 3.1 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.

3.3.3 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.

Table 3.1: *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%.

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, 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).

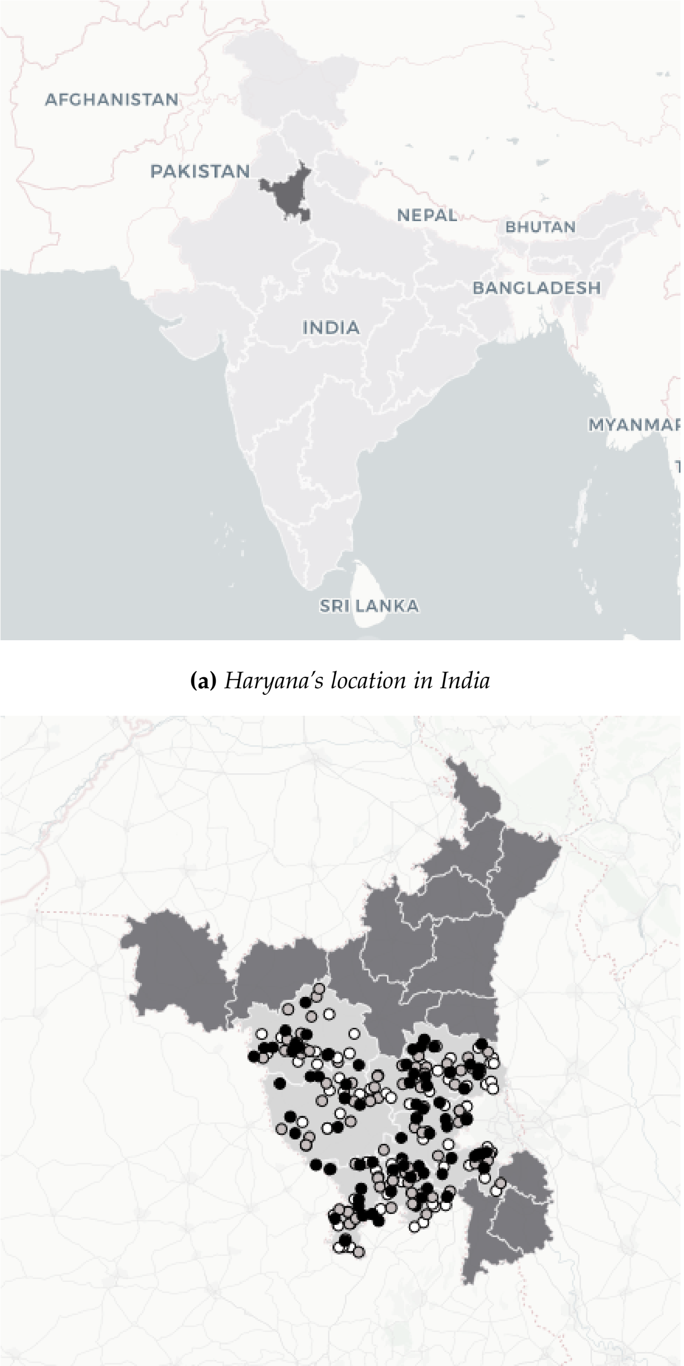
3. The **Control Group** was assigned to not receive the facilities, materials or training of the program. Its schools continued with "business-as-usual".

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.[14]

Figure 3.2 provides an overview of the study's geographic scope and the sample of schools, by treatment status.

[14] 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.* (2017b) by pre-specifing my strategy and choosing a conservative number of re-randomizations (ten re-randomizations).

Figure 3.2: *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.

3.3.4 Analytical strategy

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

$$Y_{irs2} = \alpha_s + \sum_{k=1}^{2} \beta_{ks} T_{ki} + \boldsymbol{X}_{irs1} + \phi_r + \epsilon_{irs2} \tag{3.1}$$

In Equation 3.1, 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 C.5.

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

$$Y_{irs2} = \alpha_s + \sum_{k=1}^{2} \beta_{ks} T_{ki} + \sum_{k=1}^{2} \beta_{2+ks} T_{ki} * G_{irs1} + \beta_5 G_{irs1} + \boldsymbol{X}_{irs1} + \phi_r + \epsilon_{irs2} \tag{3.2}$$

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 3.1, $\beta_{ks} \neq 0$

2. The two variants of the program affect student learning in subject s differently. H_2: in

Equation 3.1, $\beta_{1s} \neq \beta_{2s}$

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 3.2, $\beta_{2+ks} \neq 0$.

3.3.5 Balance and attrition

As shown in Table 3.2, 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 3.4. 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).[15] I address this issue with a pre-registered robustness

[15] Despite Mutz *et al.* (2018), some researchers still consider it troublesome if the treatment and control groups

Table 3.2: *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. 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%.

check, in which the most severely imbalanced triplets of schools are dropped from the statistical analysis.[16] 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 C1 show the results from this approach—by design, student performance becomes balanced across the three groups.

3.3.6 Implementation fidelity and take-up

For both treatment variants, the interventions were largely implemented as intended. As shown in Table 3.3, 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 C2, 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 3.3's Panel B, these upgrades translated into substantial usage of the

possess different means, with respect to their baseline test scores (cf. Gerber *et al.*, 2015).

[16] My work on these and other robustness checks is currently ongoing. My pre-analysis plan specifies the following checks, and their combinations: (1) Randomization inference (replicating the randomization strategy in each iteration); (2) 'exclusion' of 7 imbalanced randomization strata, via interaction effects (see above); (3) inverse-probability weights (IPW); (4) Lee (2009) bounds; (5) exclusion of repeating test items from the follow-up assessment; and (6) 'exclusion' of schools with higher cheating indices (following Jacob and Levitt, 2003, Appendix 1), via interaction effects.

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 C2 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).

3.4 Results

3.4.1 Effects on student learning

Table 3.4 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 0.15 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 0.06 standard deviations worse than Control school students, but this coefficient is not statistically significant. Workbook students performed 0.09 standard

Table 3.3: *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.

deviations better than students in ICT schools, but the results cannot reject that the difference is zero. Note how these effect sizes compare to 0.30 standard deviations of learning over the same time period, in the control group. Accordingly, the ICT intervention effectively halved students' year-to-year growth in mathematics. I do not find effects on science for either intervention.

The remaining three panels document secondary results. They provide ITT effects on student learning by cognitive domains, curricular domains, and content domain. Panel B 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. The negative effect on mathematics learning loses its statistical significance, for at-grade-level material. At the same time, for the comparison of ICT schools again Workbook schools, the coefficient for below-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.

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 five percentage points less likely to have mastered algebra, five percentage points less likely to have mastered geometry, and five percentage points less likely to have mastered the number sense domain. In a comparison with their peers in Workbook schools, I also find negative effects for biology and chemistry. The remaining coefficients are of similar magnitude, but they are not statistically distinguishable from zero.

Table 3.4: *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***	0.03	-0.15**	0.19**	-0.15**	-0.09	-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	-0.22***	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					-0.13**	-0.08*	-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					-0.02	-0.05	0.02
		[1.00]					(0.05)	(0.06)	(0.05)
Science, lower-order		0.00					-0.03	-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**	0.07**	-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	-0.06**	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	-0.08***	0.06**	-0.01	-0.03	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	-0.08***	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	0.39	0.46	0.06**	-0.00	-0.05*	0.05*	-0.05**	-0.03	-0.02
	[0.49]	[0.50]	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Geometry	0.43	0.43	0.00	0.01	-0.06**	0.07***	-0.05**	-0.04*	-0.01
	[0.50]	[0.50]	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Number sense	0.38	0.43	0.05*	-0.00	-0.05	0.04	-0.05**	-0.04*	-0.01
	[0.49]	[0.49]	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Statistics/Reasoning	0.48	0.44	-0.04*	0.00	-0.06***	0.07***	-0.03	-0.02	-0.01
	[0.50]	[0.50]	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Biology	0.47	0.46	-0.01	-0.02	-0.08***	0.06**	-0.02	-0.04**	0.02
	[0.50]	[0.50]	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Chemistry	0.51	0.47	-0.04	-0.01	-0.08***	0.07***	-0.02	-0.04*	0.02
	[0.50]	[0.50]	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Physics	0.48	0.46	-0.03	-0.03	-0.08***	0.05*	0.00	-0.03	0.03
	[0.50]	[0.50]	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)

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) (see Appendix C.4). 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 C.5). Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.

Sample. 16,612 grade-9 and grade-10 students, in the study's 240 schools.

3.4.2 Heterogeneity in treatment effects

In Table 3.5, 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 are dominated by its impact on grade-9 students. For mathematics, the difference in performance across the ICT and Control group students is 0.13 standard deviations larger for grade 9 students, in comparison to grade-10 students, in mathematics (it is 0.11 standard deviations larger in science). 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 0.14 standard deviations below that of students in Workbook schools (for science, this difference is 0.10 standard deviations, but it is not statistically significant).

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 C3 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 substantial heterogeneity in the effects on mathematics learning, for the full intervention with the ICT component—the difference in ITT effects is 0.54 standard deviations. A similar pattern emerges for science, and for the workbook-only intervention, but the differences across districts are smaller and

Table 3.5: *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***	-0.03	-0.15*	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***	0.09	-0.15*	0.24***	-0.13*	-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	0.12	0.00	0.11	0.04	0.09	-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	-0.13	-0.15*	0.02	-0.04	-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***	-0.01	-0.02	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***	0.14	-0.12	0.26**	0.03	0.06	-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]	[0.90]	(0.02)	(0.03)	(0.03)	(0.03)	(0.08)	(0.08)	(0.08)
Middle tercile	0.03	0.36	0.33***	-0.01	0.00	-0.02	-0.18**	-0.09	-0.08
	[0.26]	[0.93]	(0.02)	(0.02)	(0.02)	(0.02)	(0.07)	(0.06)	(0.06)
Top tercile	1.24	0.86	-0.38***	0.08	-0.07	0.15***	-0.10	-0.06	-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***	0.83***	-1.26***	0.04	-0.10	0.14**	0.06	0.07	-0.01
	(0.04)	(0.06)	(0.12)	(0.06)	(0.07)	(0.07)	(0.10)	(0.10)	(0.10)
Science									
Bottom tercile	-0.97	-0.25	0.72***	0.03	0.00	0.02	-0.07	-0.04	-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	-0.05	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	0.05	-0.04	0.09
	[0.56]	[0.91]	(0.02)	(0.05)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)
Top vs bottom tercile	1.99***	0.77***	-1.30***	-0.10	-0.18**	0.08	0.12	0.00	0.11
	(0.04)	(0.05)	(0.09)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
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	0.46***	-0.20	-0.14	-0.06	-0.42***	-0.25**	-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*	0.39*	0.14	0.25	0.54***	0.26	0.29
	(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	0.02
	[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	0.17	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) (see Appendix C.4). 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 C.5). Standard deviations in brackets; standard errors in parentheses (clustered at the school level). * significant at 10%; ** significant at 5%; *** significant at 1%.
Sample. 16,612 grade-9 and grade-10 students, in the study's 240 schools.

statistically indistinguishable from zero.[17]

3.4.3 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.

Instructional quality and teaching practices

The ICT program worsened the instructional quality students received. Panel A of Table 3.6 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 C3 and Appendix Table C4 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.3, 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.[18] 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 C3 and Appendix Table C4 suggest that these effects are driven by impacts in mathematics. In contrast, for the workbook-only intervention, I find negative effects in

[17] In Appendix Figure C4, 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. Yet, one district (Jhajjar) shows systematically better mathematics results, for a comparison of ICT schools with Workbook schools.

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

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 collected 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 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

Table 3.6: *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 C.5). In Panel A, "adjust." refers to an adjustment for systematic differences in comparison to video-based re-ratings (see Appendix C.3), 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.

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.[19] 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).

Student perceptions and attitudes

Both variants of the program led to negative effects on student perceptions and attitudes towards mathematics and science. Table 3.7 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.

[19] 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.

Table 3.7: *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 C.5). "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.

Appendix Table C5 repeats the above analysis by subject.[20] 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.

3.5 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.

The interpretation of results nevertheless requires some level of additional caution.

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

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.

Taken together, the results may be best interpreted as a cautionary warning that, at least in the short run, promising interventions that aim to improve instructional quality may not lead to improvements in student learning, even if they are implemented well. In the light of the Global Learning Crisis, and given the severe lack of rigorous research on how to improve teaching in developing countries, this study serves as a wake-up call for education researchers and practitioners to place additional focus on this issue.

Conclusion

In academic writing, common advice suggests first saying what you will say, saying it, and then saying what you have said. In the preceding chapters, I have adhered to this structure to the best of my abilities. What remains to be said once you have said what you have said? I conclude this dissertation with three thoughts on current developments in education research, and potential implications for how to address the ongoing learning crisis in less-developed countries.

First, a newfound wealth of data provides unprecedented opportunities for education research. For example, in Chapter 2 of this dissertation, I made use of data for the full population of Chile's public schools, rich data on all Chilean teachers, all students, and teacher-to-classroom links, covering a period of ten years. In Chapter 3, I was able to link all of India's registered schools to their geolocations, uniquely map them to their village or town, and connect these data to detailed information on socio-demographic characteristics (including from economic censuses, population censuses, and satellite-recorded night lights data). I hope that other countries follow these examples from Chile and India, and make existing data more readily accessible to researchers and practitioners. At the same time, I still see few examples of countries that successfully track the implementation of new policies, measure students' exposure to them, and systematically exploit existing data sources for policy evaluations (whether through experimental or quasi-experimental research).

Second, there is great room for improvement in the measurement of learning trajectories over time, and value in the use of more fine-grained measures of student skills. Chapter 1 of this dissertation shows how a simple, easy-to-use measure of learning led to severe underestimates of the extent of the Global Learning Crisis. Chapter 3 also shows how an intervention impacted one area of skills (in mathematics) yet left another set of skills unaffected (in science). In addition, while Chile's national assessment data allow for the tracking of student performance over time (see Chapter 2), in India, Haryana's data does not, and it is so far impossible to link Haryana's state-administered tests to government-issued student rosters (see Chapter 3). These insights call for greater investments in measures that disaggregate learning across multiple domains, and allow for the assessment of learning trajectories as students progress through school.

Third and finally, the potentially greatest challenge for research on teacher effectiveness may be its inability to detect small program effects, over short time, and especially in the secondary grades (where year-to-year student progress slows down)—even with large, well-executed experiments. A simple back-of-the-envelope calculation calls for studies to be powered well enough to detect effects of 0.05 standard deviations in student learning, for interventions that are tracked over a full school year.[21] Conservative calculations suggest that any study would thus require more than 1,775 teachers (and data on more than 21,000 of their students), to investigate a single intervention.[22] This insight may call for a reduced focus on test scores and greater investments in "surrogate" measures of teacher productivity, which can proxy for impacts on student learning. Chapter 3 shows that classroom observations may provide such a measure, but it also highlights limitations (e.g., if observations of instructional quality are systematically biased or uncorrelated with student achievement). Teacher evaluations could also serve this purpose, even if their formative use turns out to be rather limited (see Chapter 2).

[21] Consider that most labor market interventions will not lead to productivity improvements greater than .25 standard deviations. Consider also that a one-standard-deviation difference in teacher productivity corresponds to roughly 0.20 standard deviations in student learning (which corresponds to roughly one year of learning in secondary grades).

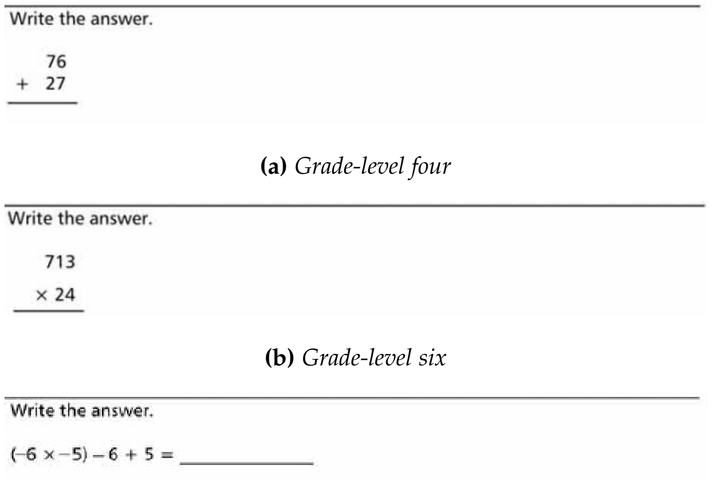
[22] This assumes the availability of a baseline assessment, a year-to-year correlation of test scores of .6, an intra-cluster correlation of .15, and the ability to sub-sample 12 students per teacher (to bring down study costs). It also (generously) assumes the ability to intervene at the teacher-level (rather than school-level), the absence of spill-over effects within schools, perfect program take-up among teachers, and that all teachers (and their students) can be tracked over a year, without attrition.

Appendix A

Supplementary Material to Chapter 1

A.1 Additional figures

Figure A1: *Examples of item classification by grade-levels: Whole number operations*



Write the answer.

$$\begin{array}{r} 76 \\ +\ 27 \\ \hline \end{array}$$

(a) *Grade-level four*

Write the answer.

$$\begin{array}{r} 713 \\ \times\ 24 \\ \hline \end{array}$$

(b) *Grade-level six*

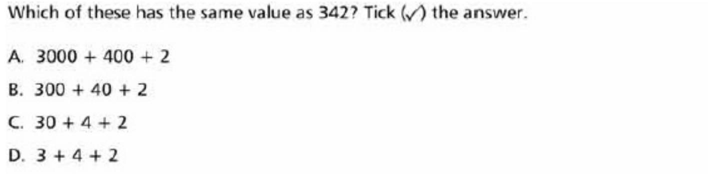
Write the answer.

$(-6 \times -5) - 6 + 5 =$ ______________

(c) *Grade-level eight*

Notes. This figure provides example items measuring the "whole number operations" skill, classified along with their respective grade-level. Panel (a) shows a fourth-grade item (item F15S9). Panel (b) shows a sixth-grade item (item S16). Panel (c) shows an eighth-grade item (item E14).

Figure A2: *Q-matrix refinement: Example item*



Which of these has the same value as 342? Tick (✓) the answer.

A. 3000 + 400 + 2

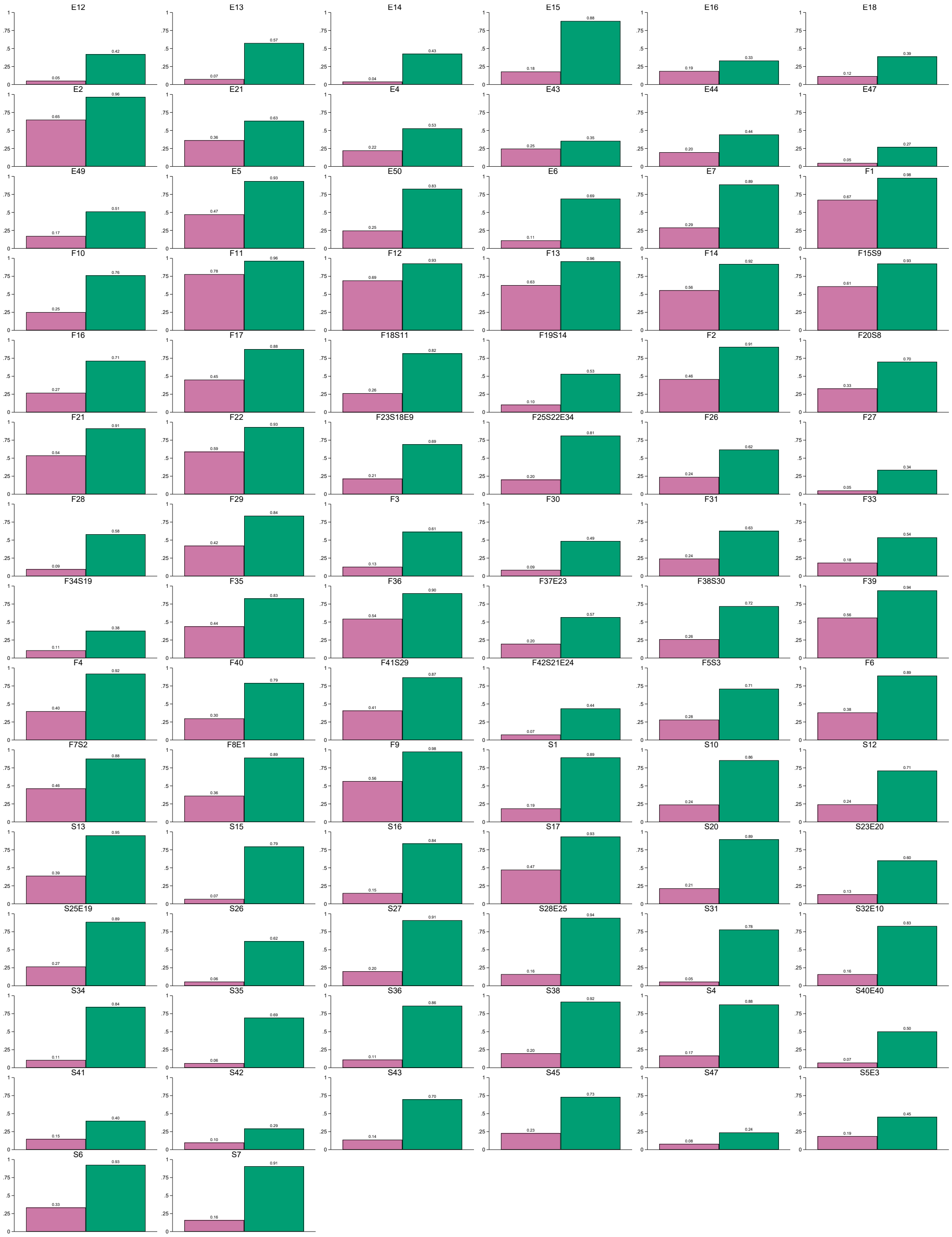
B. 300 + 40 + 2

C. 30 + 4 + 2

D. 3 + 4 + 2

Notes. This figure provides an example item (item F26) whose skill-mapping was modified as a result of a qualitative expert review, following the study's empirical Q-matrix refinement (cf. de la Torre and Chiu, 2016). In this case, number sense was removed and whole number operations was added.

Figure A3: *Item Characteristic Bar Charts (ICBCs)*



Notes. This figure provides Item Characteristic Bar Charts (ICBCs), which depict a student's expected probability of solving an item correctly, conditional on her level of mastery for the skills measured by the given item. If items measure multiple skills or grade-level expectations, we only display the expected probability for students who have not mastered any, versus students who have mastered all measured attributes.

A.2 Additional tables

Table A1: *Items, item-to-skill mapping, and item parameters*

Item	Grade level	Skill 1	Skill 2	λ_{j0}	s.e.	λ_{j1}	s.e.	λ_{j2}	s.e.	λ_{j3}	s.e.
F1	Fourth	N		0.673	0.007	0.304	0.007				
F2	Fourth	N		0.458	0.007	0.448	0.008				
F3	Fourth	N		0.128	0.005	0.486	0.007				
F4	Fourth	N		0.398	0.007	0.521	0.008				
F5S3	Fourth	N		0.281	0.005	0.429	0.006				
F6	Fourth	N		0.381	0.007	0.511	0.008				
F7S2	Fourth	N		0.463	0.005	0.413	0.006				
F8E1	Fourth	N		0.362	0.005	0.529	0.006				
F9	Fourth	G		0.565	0.006	0.410	0.006				
F10	Fourth	D		0.251	0.005	0.511	0.007				
F11	Fourth	W		0.778	0.005	0.184	0.006				
F12	Fourth	W		0.691	0.006	0.236	0.007				
F13	Fourth	W		0.626	0.006	0.330	0.007				
F14	Fourth	W		0.555	0.006	0.364	0.007				
F15S9	Fourth	W		0.610	0.004	0.315	0.005				
F16	Fourth	W		0.269	0.006	0.444	0.008				
F17	Fourth	W		0.451	0.006	0.424	0.007				
F18S11	Fourth	W		0.263	0.004	0.553	0.005				
F19S14	Fourth	W		0.104	0.003	0.427	0.005				
F20S8	Fourth	G		0.329	0.004	0.371	0.005				
F21	Fourth	M		0.536	0.006	0.376	0.007				
F22	Fourth	M		0.589	0.006	0.341	0.007				
F23S18E9	Fourth	D		0.214	0.003	0.478	0.004				
F25S22E34	Fourth	W		0.201	0.003	0.610	0.004				
F26	Fourth	W		0.238	0.005	0.380	0.008				
F27	Fourth	N		0.050	0.003	0.286	0.006				
F28	Fourth	W		0.095	0.004	0.485	0.007				
F29	Fourth	D		0.421	0.006	0.416	0.007				
F30	Fourth	M		0.085	0.003	0.401	0.007				
F31	Fourth	G		0.241	0.005	0.385	0.008				
F33	Fourth	W		0.184	0.005	0.352	0.007				
F34S19	Fourth	D		0.106	0.002	0.270	0.005				
F35	Fourth	W		0.439	0.006	0.388	0.008				
F36	Fourth	N		0.544	0.007	0.354	0.008				

Table A1 continued from previous page

Item	Grade level	Skill 1	Skill 2	λ_{j0}	s.e.	λ_{j1}	s.e.	λ_{j2}	s.e.	λ_{j3}	s.e.
F37E23	Fourth	M		0.196	0.003	0.370	0.005				
F38S30	Fourth	W		0.260	0.004	0.459	0.005				
F39	Fourth	D	W	0.559	0.007	0.342	0.013	0.328	0.010	-0.291	0.015
F40	Fourth	W		0.298	0.006	0.491	0.008				
F41S29	Fourth	W		0.407	0.004	0.461	0.005				
F42S21E24	Fourth	M		0.074	0.002	0.363	0.004				
S1	Advanced	N		0.186	0.005	0.473	0.005	0.235	0.005		
S4	Advanced	N		0.167	0.005	0.474	0.005	0.235	0.005		
S5E3	Fourth	N		0.186	0.004	0.268	0.005				
S6	Advanced	N		0.333	0.006	0.426	0.006	0.166	0.005		
S7	Advanced	N		0.157	0.005	0.485	0.005	0.264	0.005		
S10	Advanced	W		0.239	0.005	0.432	0.006	0.184	0.005		
S12	Fourth	W		0.242	0.005	0.469	0.007				
S13	Advanced	W		0.388	0.005	0.436	0.006	0.125	0.005		
S15	Advanced	W		0.067	0.003	0.287	0.006	0.440	0.006		
S16	Advanced	W		0.149	0.004	0.404	0.006	0.284	0.006		
S17	Advanced	M		0.473	0.005	0.396	0.006	0.063	0.005		
S20	Advanced	W		0.214	0.004	0.416	0.005	0.264	0.005		
S23E20	Fourth	G		0.131	0.003	0.469	0.004				
S25E19	Advanced	G		0.266	0.003	0.190	0.005	0.430	0.004		
S26	Advanced	D		0.056	0.003	0.132	0.005	0.429	0.006		
S27	Advanced	M		0.200	0.004	0.470	0.006	0.239	0.006		
S28E25	Advanced	M		0.160	0.003	0.409	0.004	0.372	0.004		
S31	Advanced	W		0.055	0.003	0.182	0.005	0.542	0.005		
S32E10	Advanced	N		0.159	0.003	0.339	0.004	0.331	0.004		
S34	Advanced	D		0.106	0.003	0.457	0.006	0.278	0.006		
S35	Advanced	W		0.062	0.003	0.163	0.006	0.468	0.006		
S36	Advanced	D		0.111	0.003	0.293	0.006	0.455	0.006		
S38	Advanced	M		0.199	0.004	0.482	0.006	0.234	0.006		
S40E40	Advanced	M		0.069	0.002	0.074	0.003	0.359	0.004		
S41	Fourth	W		0.146	0.004	0.252	0.006				
S42	Fourth	W		0.097	0.003	0.194	0.006				
S43	Advanced	D		0.137	0.003	0.189	0.006	0.371	0.006		
S45	Fourth	G		0.227	0.004	0.503	0.006				
S47	Fourth	D	W	0.078	0.003	0.010	0.011	-0.006	0.006	0.155	0.013
E2	Fourth	N		0.648	0.008	0.315	0.008				
E4	Advanced	W		0.221	0.005	0.042	0.006	0.264	0.006		
E5	Fourth	W		0.471	0.007	0.462	0.007				

Table A1 continued from previous page

Item	Grade level	Skill 1	Skill 2	λ_{j0}	s.e.	λ_{j1}	s.e.	λ_{j2}	s.e.	λ_{j3}	s.e.
E6	Advanced	W		0.110	0.004	0.175	0.005	0.404	0.005		
E7	Advanced	W		0.288	0.005	0.473	0.005	0.125	0.005		
E12	Advanced	D		0.050	0.003	0.011	0.004	0.358	0.005		
E13	Advanced	N		0.074	0.004	0.065	0.005	0.436	0.006		
E14	Advanced	W		0.039	0.003	0.026	0.004	0.361	0.006		
E15	Advanced	N		0.179	0.005	0.262	0.006	0.440	0.005		
E16	Advanced	N		0.186	0.006	-0.023	0.006	0.168	0.007		
E18	Advanced	N		0.115	0.005	-0.021	0.006	0.294	0.007		
E21	Advanced	G		0.363	0.005	0.189	0.006	0.080	0.006		
E43	Advanced	W		0.246	0.005	0.057	0.006	0.051	0.006		
E44	Advanced	M		0.196	0.005	0.102	0.006	0.145	0.006		
E47	Advanced	D		0.047	0.003	0.021	0.004	0.201	0.005		
E49	Advanced	G		0.171	0.004	0.127	0.006	0.213	0.006		
E50	Advanced	W		0.246	0.005	0.414	0.005	0.165	0.005		

Notes. Item names reflect the grade in which items were administered and the question number in each of the three instruments. F denotes fourth-, S denotes sixth-, E denotes eighth-grade students. “Fourth” indicates material up to grade four; “Advanced” indicates material beyond grade four. Skills are denoted as follows: Decimals and fractions (D); shapes and geometry (G); measurement (M); number sense (N); whole number operations (W).

Table A2: *Mastery of fourth-grade skills, by students' enrolled grade-level and state*

State	Class	Fractions and Decimals	Measurement	Number Concepts	Operations on Whole Numbers	Shapes and Geometry
Andhra Pradesh	4	0.474	0.271	0.631	0.403	0.497
Andhra Pradesh	6	0.373	0.375	0.635	0.480	0.417
Andhra Pradesh	8	0.504	0.431	0.663	0.637	0.379
Assam	4	0.485	0.345	0.392	0.389	0.437
Assam	6	0.536	0.362	0.611	0.558	0.555
Assam	8	0.373	0.491	0.762	0.692	0.406
Bihar	4	0.424	0.278	0.547	0.517	0.432
Bihar	6	0.528	0.480	0.692	0.731	0.539
Bihar	8	0.500	0.567	0.771	0.825	0.569
Chandigarh	4	0.530	0.637	0.838	0.448	0.560
Chandigarh	6	0.595	0.671	0.729	0.668	0.753
Chandigarh	8	0.573	0.623	0.908	0.966	0.836
Chhattisgarh	4	0.484	0.283	0.542	0.415	0.514
Chhattisgarh	6	0.330	0.294	0.696	0.451	0.547
Chhattisgarh	8	0.375	0.426	0.716	0.536	0.581
Delhi	4	0.718	0.528	0.832	0.751	0.779
Gujarat	4	0.496	0.259	0.503	0.411	0.437
Gujarat	6	0.402	0.299	0.585	0.457	0.473
Gujarat	8	0.522	0.481	0.719	0.595	0.546
Haryana	4	0.500	0.306	0.500	0.490	0.492
Haryana	6	0.663	0.374	0.722	0.708	0.634
Haryana	8	0.627	0.653	0.813	0.900	0.785
Jammu and Kashmir	4	0.150	0.132	0.154	0.088	0.203
Jammu and Kashmir	6	0.236	0.166	0.172	0.221	0.263
Jammu and Kashmir	8	0.393	0.179	0.227	0.516	0.262
Jharkhand	4	0.441	0.403	0.519	0.439	0.487
Jharkhand	6	0.498	0.548	0.669	0.669	0.606
Jharkhand	8	0.564	0.558	0.692	0.783	0.602
Karnataka	4	0.701	0.652	0.734	0.568	0.645
Karnataka	6	0.859	0.836	0.868	0.806	0.740
Karnataka	8	0.626	0.748	0.838	0.793	0.556
Kerala	4	0.587	0.684	0.907	0.559	0.740
Kerala	6	0.845	0.661	0.631	0.785	0.681
Kerala	8	0.810	0.857	0.943	0.885	0.721
Madhya Pradesh	4	0.339	0.255	0.412	0.290	0.369
Madhya Pradesh	6	0.316	0.248	0.488	0.396	0.461
Madhya Pradesh	8	0.457	0.419	0.636	0.616	0.567
Maharashtra	4	0.623	0.678	0.833	0.765	0.641
Maharashtra	6	0.725	0.787	0.870	0.841	0.828
Maharashtra	8	0.695	0.692	0.807	0.757	0.786
Odisha	4	0.585	0.497	0.755	0.562	0.642
Odisha	6	0.557	0.674	0.756	0.661	0.687
Odisha	8	0.579	0.548	0.858	0.813	0.605
Punjab	4	0.412	0.637	0.843	0.227	0.528
Punjab	6	0.458	0.690	0.588	0.565	0.438
Punjab	8	0.840	0.706	0.869	0.867	0.453
Rajasthan	4	0.393	0.326	0.344	0.317	0.453
Rajasthan	6	0.389	0.283	0.501	0.365	0.455
Rajasthan	8	0.600	0.635	0.734	0.673	0.666
Tamil Nadu	4	0.562	0.388	0.793	0.446	0.547
Tamil Nadu	6	0.708	0.548	0.742	0.614	0.443
Tamil Nadu	8	0.727	0.599	0.732	0.653	0.520
Uttarakhand	4	0.579	0.460	0.667	0.528	0.618
Uttarakhand	6	0.498	0.497	0.715	0.600	0.659
Uttarakhand	8	0.706	0.621	0.795	0.720	0.679

Notes. This table reports on the probability that a student has mastered fourth-grade material, by students' enrolled grade-level, state, and skill. In Delhi, only fourth-graders were tested.

Table A3: *Gender skill gaps, by students' enrolled grade-level and state*

State	Class	Fractions and Decimals	Measurement	Number Concepts	Operations on Whole Numbers	Shapes and Geometry
Andhra Pradesh	4	0.001	0.079	0.028	0.047	0.002
Andhra Pradesh	6	0.014	0.107	0.083	-0.023	0.058
Andhra Pradesh	8	0.066	0.115	0.087	0.001	0.020
Assam	4	-0.032	0.015	0.129	0.013	-0.051
Assam	6	0.063	0.132	0.115	0.120	-0.009
Assam	8	0.038	0.157	0.135	0.215	0.081
Bihar	4	0.151	0.073	0.176	0.204	0.118
Bihar	6	0.135	0.246	0.194	0.138	0.131
Bihar	8	0.182	0.225	0.191	0.142	0.117
Chandigarh	4	0.004	-0.017	0.079	0.107	-0.047
Chandigarh	6	-0.065	0.040	0.004	0.003	0.050
Chandigarh	8	-0.040	0.166	0.069	0.037	-0.041
Chhattisgarh	4	-0.049	0.027	0.007	-0.015	-0.082
Chhattisgarh	6	0.065	0.198	0.156	0.146	0.138
Chhattisgarh	8	0.109	0.149	0.153	0.131	0.083
Delhi	4	0.105	0.224	0.097	0.171	0.049
Gujarat	4	-0.018	-0.028	0.053	0.049	-0.017
Gujarat	6	0.133	0.087	0.132	0.142	0.119
Gujarat	8	0.047	0.026	0.028	0.061	-0.021
Haryana	4	0.071	-0.026	0.030	0.058	0.007
Haryana	6	-0.037	0.131	0.068	-0.013	0.029
Haryana	8	-0.003	0.149	0.127	0.018	0.031
Jammu and Kashmir	4	-0.090	-0.014	-0.064	-0.039	-0.049
Jammu and Kashmir	6	0.060	0.054	0.025	0.047	0.006
Jammu and Kashmir	8	-0.032	0.108	-0.016	0.098	-0.018
Jharkhand	4	-0.023	0.068	0.135	0.197	0.086
Jharkhand	6	0.113	0.203	0.211	0.217	0.267
Jharkhand	8	0.070	0.259	0.227	0.141	0.134
Karnataka	4	-0.045	-0.063	-0.051	-0.045	-0.097
Karnataka	6	0.016	0.072	-0.005	-0.028	-0.049
Karnataka	8	-0.031	-0.007	0.044	0.039	0.015
Kerala	4	-0.039	-0.034	-0.012	0.053	-0.065
Kerala	6	-0.031	0.021	-0.010	0.066	-0.010
Kerala	8	0.027	0.059	0.003	-0.016	-0.073
Madhya Pradesh	4	0.096	0.073	0.096	0.094	0.076
Madhya Pradesh	6	0.099	0.058	0.102	0.064	0.023
Madhya Pradesh	8	0.059	0.019	0.131	0.043	0.070
Maharashtra	4	-0.036	-0.044	0.005	-0.027	-0.046
Maharashtra	6	0.083	0.068	0.029	0.038	0.041
Maharashtra	8	0.005	0.044	0.036	-0.003	0.048
Odisha	4	0.009	-0.010	-0.003	0.015	0.009
Odisha	6	-0.015	0.018	0.006	-0.064	-0.004
Odisha	8	-0.007	0.061	-0.008	0.002	-0.039
Punjab	4	-0.062	0.023	0.015	0.014	0.008
Punjab	6	-0.053	0.022	0.009	0.025	-0.073
Punjab	8	0.008	0.024	-0.053	-0.044	-0.012
Rajasthan	4	0.066	0.081	0.112	0.053	0.018
Rajasthan	6	0.005	0.091	0.146	0.044	0.060
Rajasthan	8	0.052	0.054	0.020	0.038	-0.007
Tamil Nadu	4	-0.020	0.013	-0.038	0.016	-0.060
Tamil Nadu	6	0.041	0.134	0.086	0.033	-0.044
Tamil Nadu	8	0.004	0.092	0.072	0.081	0.004
Uttarakhand	4	0.063	0.101	0.084	0.079	0.096
Uttarakhand	6	0.003	-0.053	0.127	0.012	0.090
Uttarakhand	8	-0.002	0.073	0.034	0.123	0.056

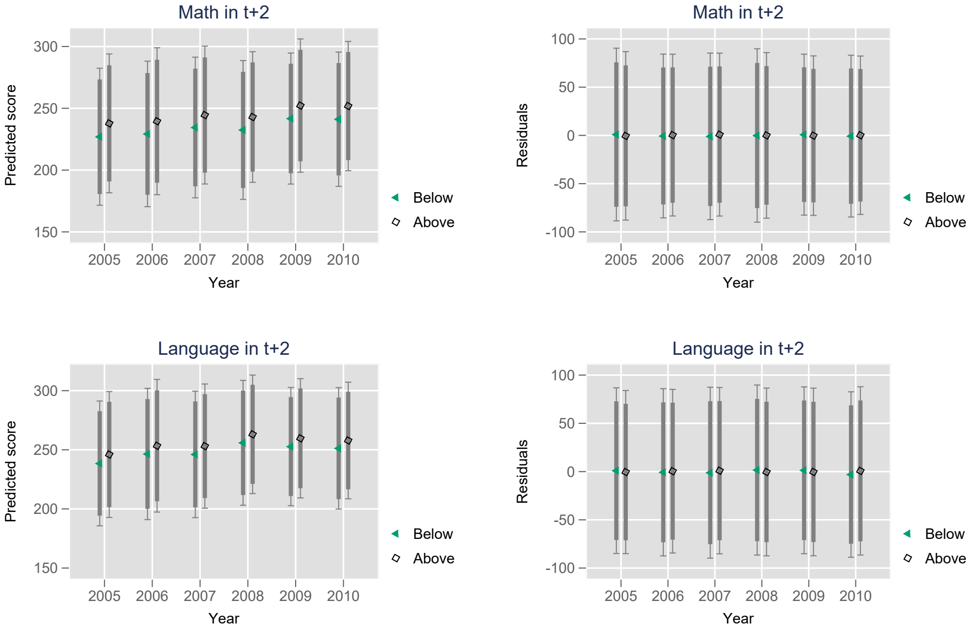
Notes. This table reports on gender gaps in the mastery of fourth-grade material, by students' enrolled grade-level, state, and skill. "Gender gap" refers to the probability of male students who have mastered a skill, minus the respective probability for female students. In Delhi, only fourth-graders were tested.

Appendix B

Supplementary Material to Chapter 2

B.1 Additional figures

Figure B1: *(Absence of) differential trends during the pre-policy period*



Notes. This figure investigates differential trends in test-scores, during the pre-policy period. Left panels show predicted scores; right panels show residuals. Predicted scores and residuals stem from a reduced form regression, as shown in Equation 2.1. Top panels refer to math outcomes in year $t + 2$; bottom panels refer to language outcomes in year $t + 2$. “Below" refers to teachers qualifying for a “basic” rating (in year t); “above” refers to teachers qualifying for a better rating (in year t).

B.2 Additional tables

Table B1: *Additional sample characteristics and validity checks*

	2005-2010		2011-13		
	Below	Above	Below	Above	DD
Attrition					
Teacher Baseline Characteristics					
t+1: Gender: Female	0.72	0.8	0.78	0.86	0.01 (0.03)
t+1: Contract hours	39.03	38.49	37.92	37.46	-0.12 (0.50)
t+1: Works in yet another school	0.08	0.07	0.04	0.02	0.01 (0.02)
t+1: Years in service	20.29	19.29	14.39	14.55	-0.53 (0.76)
t+1: n	1,972	3,700	296	1,665	7,633
t+1: School's baseline reading score†	241.02	247.14	248.88	253.48	-0.84 (1.50)
t+1: School's baseline math score†	229.19	235.7	237.88	244.72	-0.61 (1.72)
t+1: n	1,439	2,953	243	1,526	6,161
t+3: Gender: Female	0.75	0.83	0.85	0.88	0.04 (0.03)
t+3: Contract hours	38.67	38.24	36.8	36.69	-0.07 (0.63)
t+3: Works in yet another school	0.06	0.07	0.03	0.03	0.01 (0.02)
t+3: Years in service	20.12	18.42	12.16	13.89	-2.44 (0.95)**
t+3: n	1,795	3,402	185	795	6,177
t+3: School's baseline reading score†	241.15	247.64	251.75	255.94	1.40 (1.79)
t+3: School's baseline math score†	229.89	235.95	242.43	247.84	1.29 (2.06)
t+3: n	1,330	2,786	171	764	5,051
Student Baseline Characteristics					
t+1: GPA	5.77	5.83	5.73	5.79	0.01 (0.02)
t+1: Repeated in baseline year	0.03	0.02	0.02	0.02	-0.00 (0.00)
t+1: Attendance	92.78	93.12	90.99	91.9	-0.34 (0.29)
t+1: Gender: Female	0.48	0.49	0.48	0.49	-0.01 (0.01)
t+1: n	39,256	81,126	5,429	36,330	162,141
t+1: Household income (pesos)††	264,640	267,643	301,820	307,547	-7,044.85 (9,595.34)
t+1: Mother's edu. (years)††	9.84	10.02	10.4	10.88	-0.11 (0.12)
t+1: n	33,873	70,040	4,670	32,061	140,644
t+3: GPA	6.01	6.06	5.96	6	0.00 (0.03)
t+3: Repeated in baseline year	0.04	0.03	0.04	0.04	-0.00 (0.00)
t+3: Attendance	92.24	92.46	90.26	91.06	-0.11 (0.30)
t+3: Gender: Female	0.48	0.48	0.48	0.49	-0.01 (0.01)
t+3: n	35,051	74,699	3,998	18,817	132,565
t+3: Household income (pesos)††	279,234.67	288,412	358,775.1	339,968.9	16,032.12 (18,434.46)
t+3: Mother's edu. (years)††	10	10.22	11.36	11.36	0.01 (0.15)
t+3: n	30,140	64,254	3,459	16,600	114,453

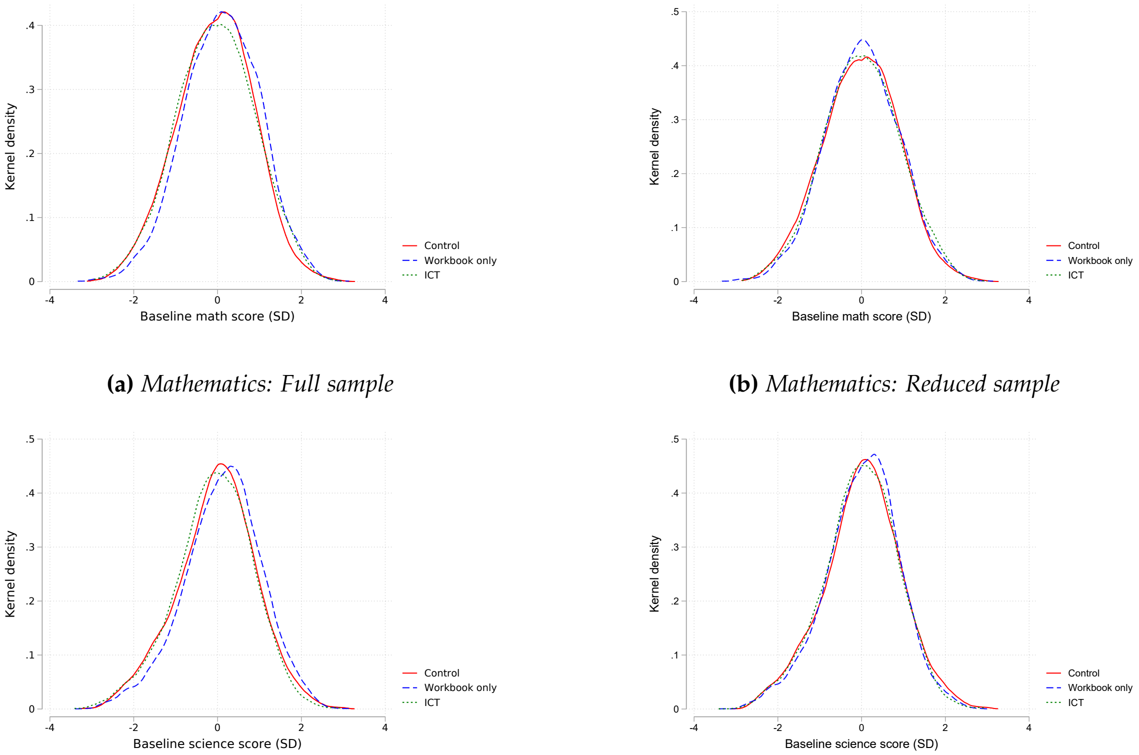
Notes. "Teachers" include all unique year-teacher observations and may thus repeatedly include individual teachers over time. "Students" include all unique year-teacher-student observations and may thus include up to two observations per student and year (if math and reading are taught by different teachers, in a given year). "Below" and "Above" refer to teachers below or above the cut-off, respectively. t refers to the year of the initial evaluation. All variables measured in t, if not denoted otherwise. † denotes variables available for fewer observations (and not included as covariates). †† denotes variables measured at follow-up (and not included as covariates). Note that the 2013 sample is not followed up in $t+3$. "DD" refers to a difference-in-difference estimate as described in Section 2.4 (excluding control variables but including commune-level fixed effects). Standard errors in parentheses. For student-level characteristics, standard errors are clustered at the year-teacher level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix C

Supplementary Material to Chapter 3

C.1 Additional figures

Figure C1: *Balance on baseline test scores.*



(a) *Mathematics: Full sample*

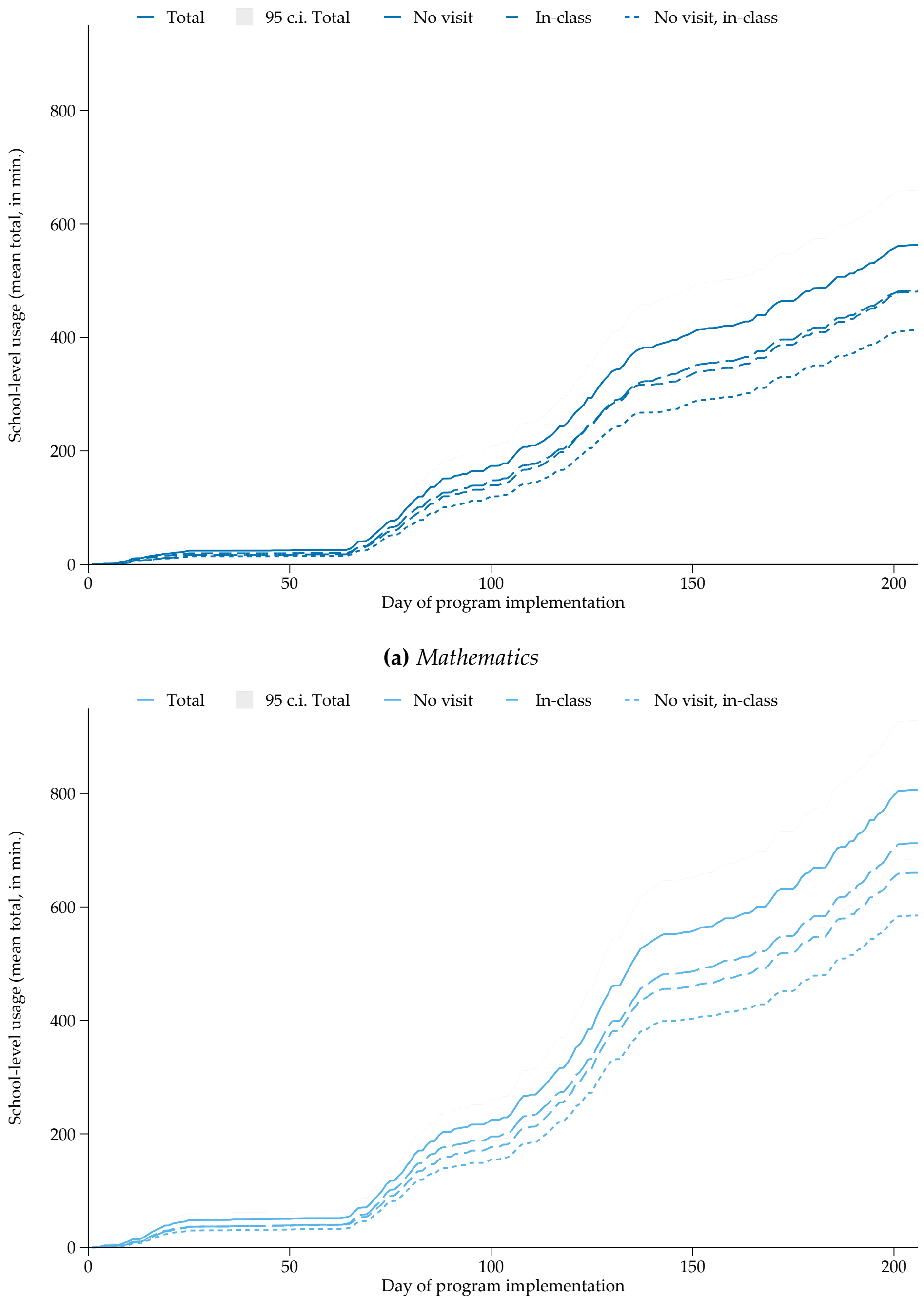
(b) *Mathematics: Reduced sample*

(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 C2: *Cumulative software usage in ICT schools, over time*

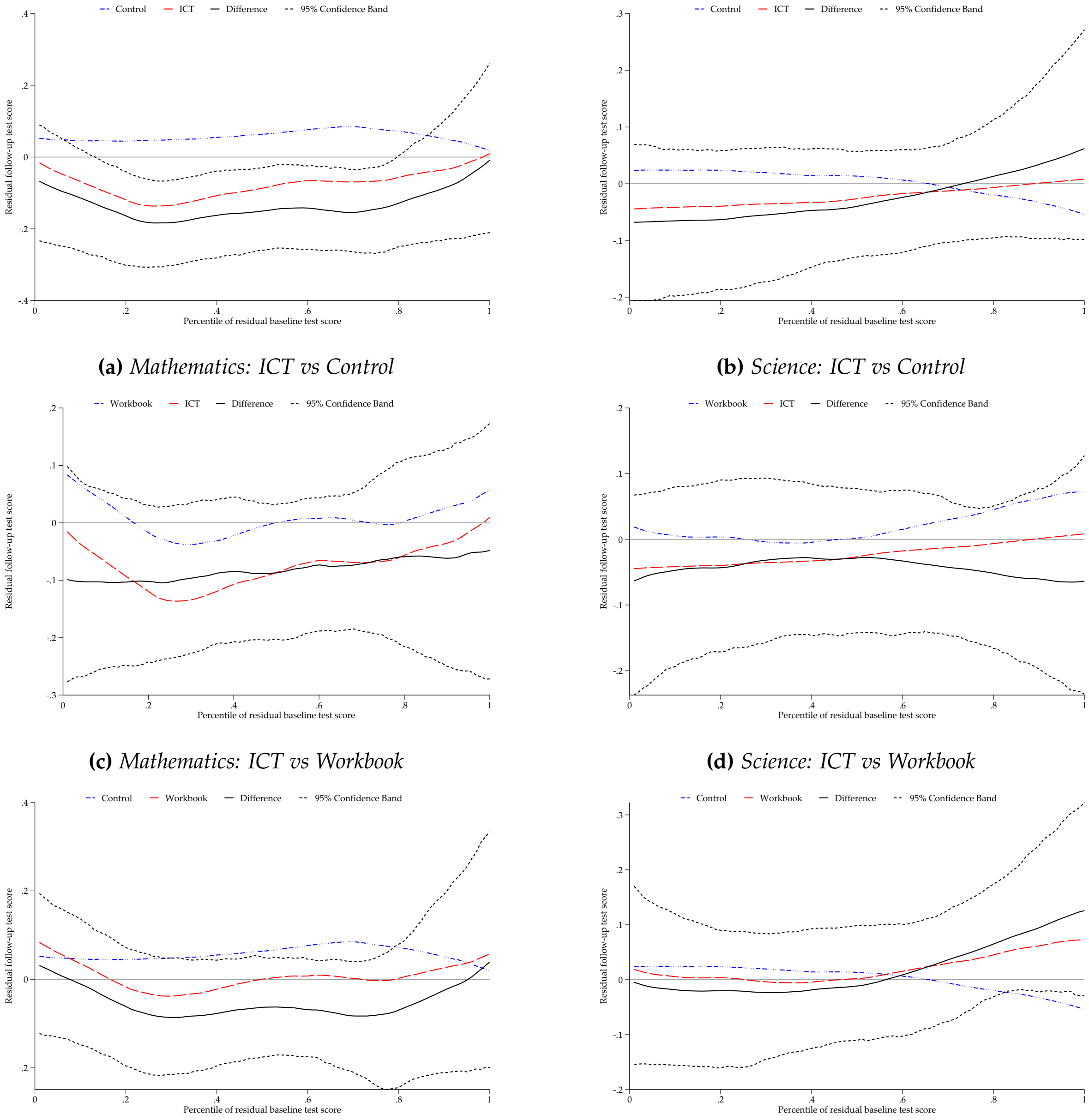


(a) *Mathematics*

(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 C3: *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*

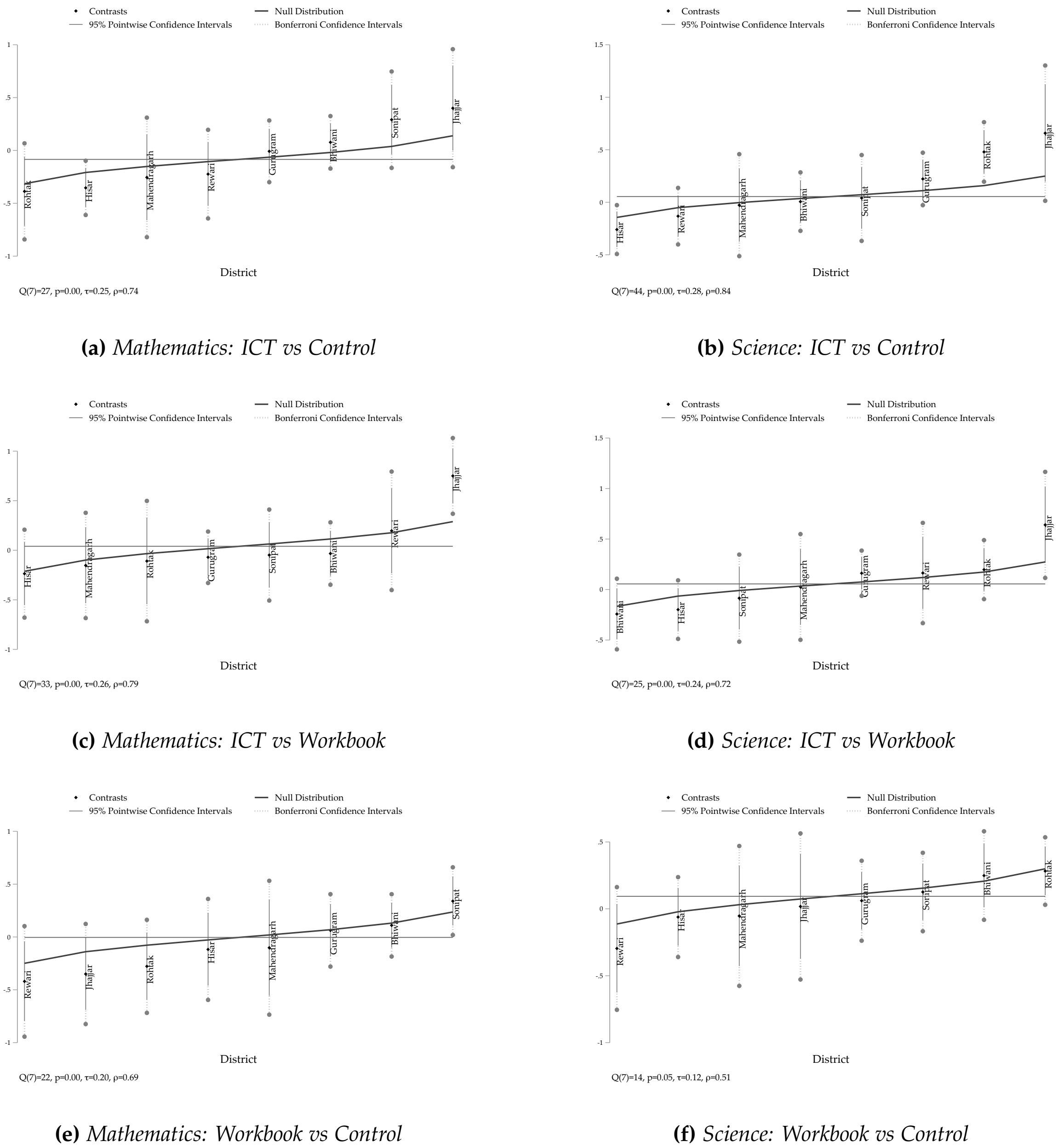
(d) *Science: ICT vs Workbook*

(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 student's test score at baseline (residualized, to account for randomization strata fixed effects). 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 C.5). "Baseline" refers to the assessment conducted in December 2018. "Follow-up" refers to the assessment conducted in November 2019.

Figure C4: *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 C.5). 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).

C.2 Additional tables

Table C1: *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 C2: *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 C3: *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 C.5). In Panel A, “adjust.” refers to an adjustment for systematic differences in comparison to video-based re-ratings (see Appendix C.3), 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 C4: *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 C.5). In Panel A, “adjust.” refers to an adjustment for systematic differences in comparison to video-based re-ratings (see Appendix C.3), 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 C5: *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 C.5). "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.3 Measuring instructional quality

C.3.1 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.*, 2018, 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.3.2 Content

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.

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.3.3 Administration

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.[1] 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.

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-

[1] 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.

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.[2] 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.3.4 Scoring

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.

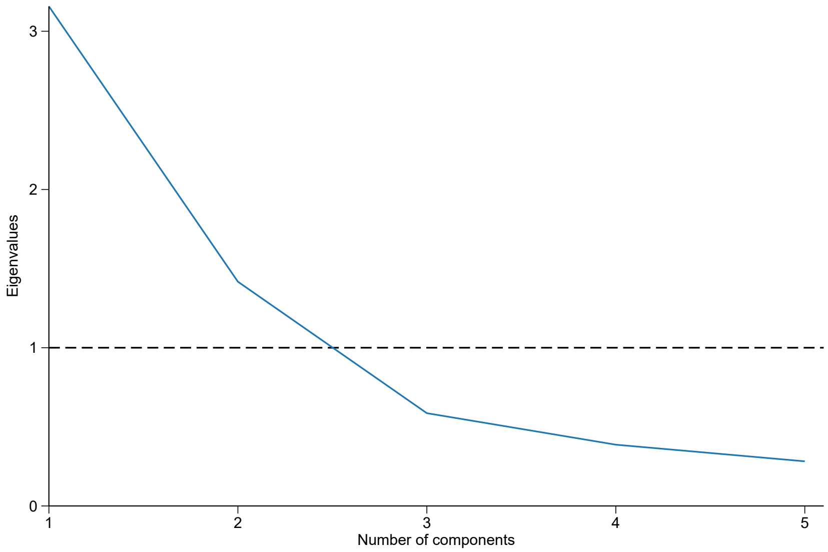
Alternative approach, dimensionality

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.*, 2018). 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 C5, 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

[2] 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 C5: *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.

second.

Similarly, in Figure C6, 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.

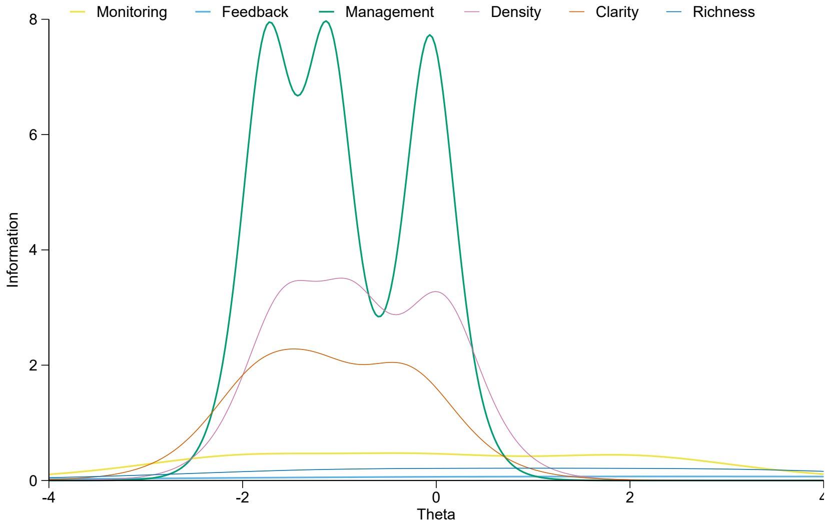
C.3.5 Empirical distribution of scores

Figure C7 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.

Second, however, this is contrasted by lower scores along dimensions that tap into

Figure C6: *Item information curves for a QUIP index that uses IRT*



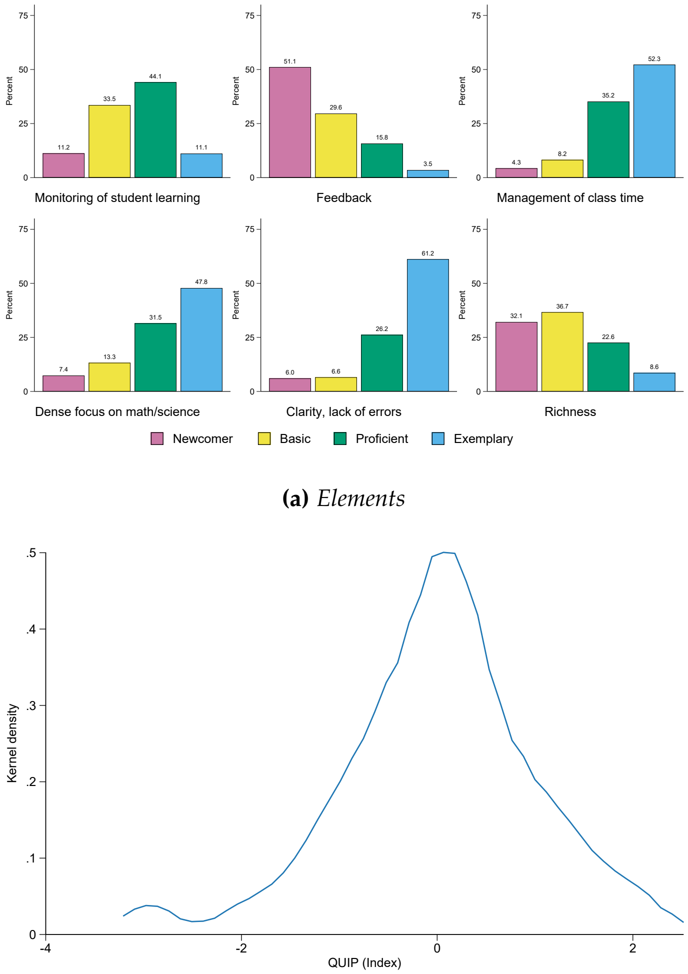
Notes. This figure provides item information functions as estimated from a Graded Response Model.

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.

Figure C7: *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.

C.3.6 Coherence

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.

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 C6 shows the results from this analysis.

Table C6: *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.

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 C7.

Table C7: *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%.

In Table C7, 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.

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 C7 (Columns 3 to 6) from each snippet's in-person rating, prior to calculating the standardized QUIP scores and their index.

C.3.7 Predictive associations

Finally, in Table C8, 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 C8: *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 C.5). Standard errors in parentheses, clustered at the school level. * significant at 10%; ** significant at 5%; *** significant at 1%.

C.4 Measuring student learning

C.4.1 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.

C.4.2 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 3.3.1, I have already given a broad overview of these domains. Here, in Table C9, 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

Table C9: *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)).

(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.

C.4.3 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 C9 summarizes the resulting number of repeated items across grades (Column 11) and test occasions (Column 12).

C.4.4 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.

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.[3]

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.

[3] 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).

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.

C.4.5 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.

Continuous scoring using Item Response Theory (IRT)

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.[4] In this study, I calculate scaled scores with a standard, two-parameter logistic (2PL) IRT model, which explicitly models each question's difficulty and its ability to discriminate (Birnbaum, 1968; Samejima, 1973).[5]

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 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.[6]

Determining student proficiency using Cognitive Diagnostic Models (CDMs)

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 (de la Torre, 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.

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

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

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

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 de la Torre (2011), we may express a respondent's probability of solving an item as

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

, 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.[7]

C.4.6 Empirical distribution of scores

In Figure C8, I show kernel density plots for the empirical distribution of test scores, for students in the Control group.[8] 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

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

[8] The figure complements Figure C1, which shows kernel density plots at baseline, across the three experimental groups.

for) the tests' precision, including for students of very low (/very high) ability.

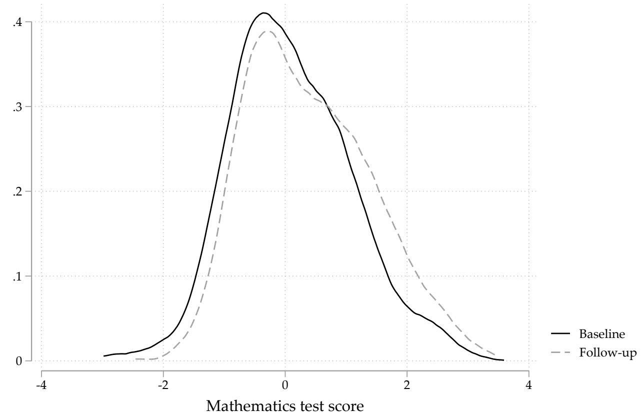
C.4.7 Item fit and test coherence

Table C10 and Table C11 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.

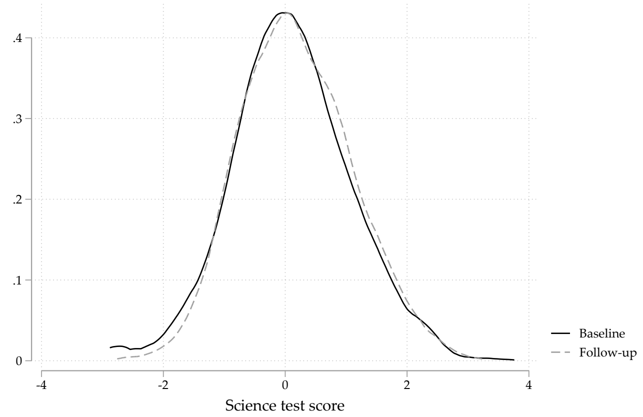
Combined with the test length, these item characteristics translate into high levels of internal consistency. One benefit of item response theory is the ability to report on test precision across a *range* of student ability, not just a single measure of test reliability (such as Cronbach's alpha, for 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 C9 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.

Figure C8: *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 C10: *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	-0.32	0.57			
P027952	7	n/a	0.64	-0.53	0.57	0.55		
P027972	8	n/a	0.91	0.53	0.39			
P051137	7	n/a	0.96	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	1.54	-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	1.48	0.09	0.46		0.47	0.52
P085375	9	Lower-order	1.68	0.01		0.46	0.48	
P085378	9	Lower-order	1.28	0.67		0.32	0.33	
P085521	8	Lower-order	1.09	-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	1.08	0.46				0.44
P087694	10	Lower-order	1.30	0.77				0.36
P103034	8	Higher-order	0.89	1.00			0.32	
P104677	10	Higher-order	0.62	1.28				0.35
P107067	10	Higher-order	0.64	0.55				0.45
P107070	9	Lower-order	0.88	0.13			0.48	
Geometry								
P021487	6	n/a	1.11	-1.55	0.79			
P021519	6	n/a	0.71	-0.02	0.49			
P027794	7	n/a	1.48	-0.74	0.64	0.67		
P027808	7	n/a	0.34	3.05	0.27	0.26		
P027948	9	n/a	0.93	-0.45		0.56		
P027955	8	n/a	0.74	0.38	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	1.07	0.16	0.46	0.41	0.46	0.55
P085373	9	Lower-order	1.42	-0.50		0.57	0.65	
P085384	9	Lower-order	1.09	-0.08		0.50	0.50	
P085416	8	Lower-order	1.00	0.93		0.30	0.30	
P085502	7	Higher-order	0.86	1.29	0.28			0.29
P085562	8	Higher-order	0.94	-0.26	0.58			0.54
P048385	10	Lower-order	0.12	-0.08				0.51
P087395	9	Lower-order	1.25	-0.03			0.51	
P087698	10	Higher-order	0.17	7.72				0.22
P087699	10	Lower-order	0.63	2.12				0.25
P087702	10	Higher-order	0.54	2.24				0.26
P103017	7	Lower-order	0.96	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	1.66	0.31			
P084830	8	Lower-order	1.36	-0.35	0.57	0.53	0.64	0.62
P084888	8	Lower-order	0.85	0.89	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	1.36	0.49	0.37			0.41
P085546	7	Lower-order	1.20	0.42	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	1.09	-0.09				0.56
P104334	7	Lower-order	1.09	0.52			0.39	
P104674	10	Lower-order	1.01	0.60				0.42
P104675	10	Higher-order	1.16	1.07				0.31
P104684	10	Higher-order	0.85	0.08				0.52
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	2.43	0.25	0.28		
P043813	9	n/a	1.04	-0.09		0.49		
P051127	7	n/a	0.83	-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	0.87	-0.24	0.52			0.60
P084836	7	Lower-order	1.28	-0.79	0.66	0.62	0.70	0.77
P084891	8	Higher-order	0.81	0.81	0.38	0.35		0.34
P084893	8	Higher-order	1.26	0.60	0.33		0.37	0.40
P085393	9	Higher-order	0.69	1.92		0.21	0.25	
P085413	9	Higher-order	0.72	1.01		0.33	0.34	
P066086	10	Lower-order	1.44	-0.06				0.56
P087405	8	Higher-order	0.76	-0.20			0.54	
P087558	8	Higher-order	0.76	0.48			0.42	
P087717	10	Higher-order	0.62	0.84				0.41
P087719	10	Higher-order	1.24	0.36				0.46
P087722	10	Lower-order	0.82	0.80				0.40
P103021	9	Lower-order	0.75	0.74			0.38	
P107071	9	Higher-order	0.26	4.95			0.22	

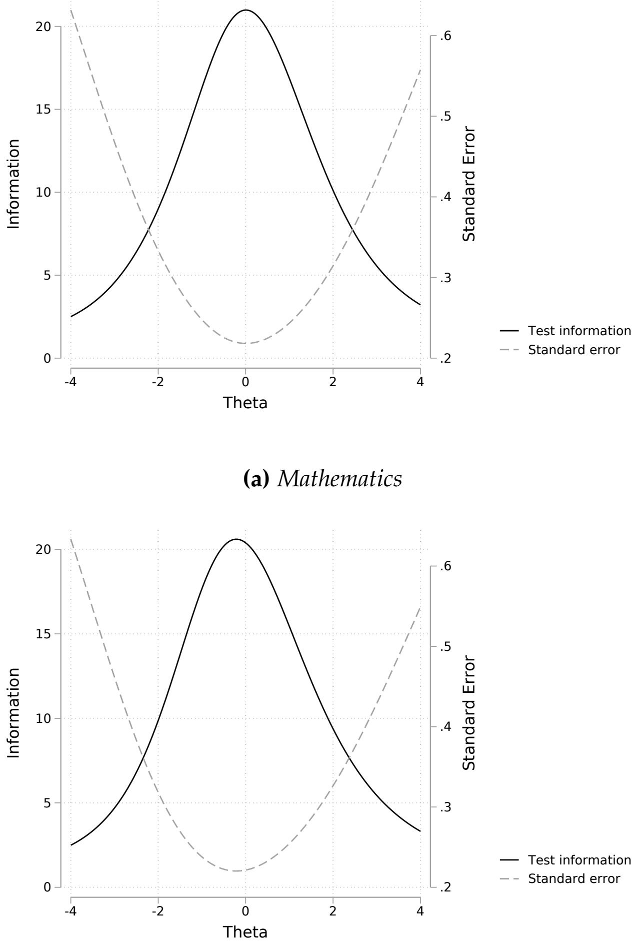
Notes. This table provides item characteristics as per a 2PL item response theory (IRT) model. Items are sorted by content domain. Item names refer to study-internal question IDs. For reference, the table also provides each items' grade-level mapping (Column 1), whether the item is mapped to higher- 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 C11: *Item characteristics (IRT): Science*

	Mapping		Item parameters (IRT 2PL)		Percent correct			
	Grade-level	Higher/lower	Discrimination	Difficulty	Baseline 9	Baseline 10	Follow-up 9	Follow-up 10
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Biology								
P021356	9	n/a	1.00	0.20		0.43		
P021631	6	n/a	1.07	-0.08	0.50	0.51		
P021634	6	n/a	1.21	0.28	0.44			
P021643	6	n/a	1.06	-0.04	0.51			
P022088	9	n/a	0.70	0.34		0.43		
P026003	9	n/a	0.72	0.65		0.38		
P026025	8	n/a	0.81	-0.02	0.51	0.48		
P054655	9	n/a	0.77	1.15		0.30		
P084959	7	Lower-order	0.99	1.23	0.28			0.27
P084961	8	Lower-order	1.11	-0.16	0.52	0.50	0.53	0.59
P084963	8	Lower-order	1.59	-0.55	0.64	0.60	0.65	0.71
P084973	8	Lower-order	0.76	0.85	0.35	0.34	0.38	
P085428	9	Lower-order	0.94	-0.23		0.52	0.56	
P085429	9	Lower-order	1.37	-0.90		0.68	0.74	
P085431	9	Lower-order	0.64	1.56		0.27	0.28	
P085536	8	Higher-order	0.91	0.54	0.45	0.36		0.39
P085537	8	Higher-order	0.74	1.42	0.27	0.28		0.29
P085538	7	Lower-order	0.84	0.40	0.45			0.43
P051367	10	Lower-order	1.35	-0.13				0.56
P086914	8	Lower-order	1.11	-0.14			0.53	
P087469	9	Lower-order	0.80	0.05			0.49	
P087471	7	Lower-order	1.22	-0.07			0.52	
P087475	9	Higher-order	0.12	10.72			0.22	
P087659	10	Higher-order	1.09	0.39				0.44
P087661	10	Lower-order	1.55	-0.92				0.76
P087663	10	Higher-order	0.94	-0.38				0.60
P087666	10	Higher-order	1.08	-0.50				0.63
P087667	10	Higher-order	1.11	0.07				0.51
P103037	9	Lower-order	0.38	1.72			0.35	
P103046	7	Higher-order	1.08	-0.36			0.58	
Chemistry								
P021392	9	n/a	0.69	1.98		0.21		
P021399	9	n/a	0.89	0.65		0.36		
P021401	9	n/a	0.89	0.87		0.32		
P021628	7	n/a	0.72	0.36	0.45			
P025446	8	n/a	0.93	0.26	0.48	0.40		
P025450	8	n/a	1.31	-0.44	0.60	0.58		
P025452	8	n/a	0.51	0.34	0.48	0.43		
P039159	8	n/a	1.45	-0.13	0.54			
P056305	7	n/a	1.01	0.08	0.49			
P072312	9	n/a	0.56	2.39		0.21		
P084976	8	Higher-order	1.18	0.37	0.42	0.36	0.40	0.50
P084986	8	Lower-order	1.39	-1.16	0.77	0.74	0.76	0.81
P084991	8	Higher-order	1.43	-0.51	0.62	0.58	0.64	0.70
P085044	7	Higher-order	1.01	-0.03	0.53			0.50
P085433	9	Lower-order	1.41	-1.04		0.72	0.76	
P085435	9	Lower-order	1.13	-0.07		0.51	0.49	
P085437	9	Higher-order	1.12	-0.03		0.49	0.49	
P085539	8	Lower-order	1.06	0.19	0.43			0.51
P085540	8	Higher-order	0.66	1.37	0.33			0.29
P086928	8	Higher-order	1.31	0.44			0.39	
P086932	7	Higher-order	0.90	0.74			0.36	
P087409	9	Lower-order	0.64	1.18			0.34	
P087410	9	Higher-order	1.51	-0.62			0.67	
P087412	8	Lower-order	1.55	-0.37			0.60	
P087676	10	Higher-order	1.02	0.54				0.41
P087677	10	Higher-order	1.20	-0.20				0.57
P087678	10	Lower-order	0.99	-0.52				0.63
P087680	10	Lower-order	0.62	0.17				0.49
P087681	10	Higher-order	1.13	0.18				0.48
P087683	10	Lower-order	0.79	1.05				0.34
P103043	9	Lower-order	0.33	0.97			0.42	
Physics		n/a						
P013841	7	n/a	0.75	-0.01	0.53	0.46		
P021822	9	n/a	0.54	1.38		0.32		
P024152	8	n/a	0.66	-0.16	0.53	0.50		
P024908	8	n/a	0.53	1.64	0.30	0.30		
P025624	8	n/a	0.57	0.77	0.40			
P054668	8	n/a	0.76	0.07	0.44	0.51		
P054831	7	n/a	0.53	1.16	0.35	0.36		
P059488	9	n/a	0.74	1.13		0.31		
P084898	7	Higher-order	1.32	-1.07	0.74	0.71	0.78	0.77
P084900	8	Higher-order	0.80	0.05	0.50	0.46	0.50	0.51
P084906	8	Lower-order	0.75	-0.31	0.59	0.53	0.57	0.52
P084953	8	Lower-order	1.29	-0.51	0.62			0.66
P085419	9	Lower-order	1.00	0.16		0.43	0.48	
P085426	9	Higher-order	0.89	0.06		0.48	0.48	
P085427	9	Lower-order	0.56	0.98		0.36	0.38	
P085534	7	Lower-order	1.12	-0.00	0.51			0.52
P085535	8	Higher-order	0.55	1.57	0.35			0.28
P087419	8	Lower-order	1.23	-0.43			0.60	
P087420	8	Lower-order	0.45	0.58			0.44	
P087668	10	Lower-order	0.74	0.87				0.38
P087670	10	Lower-order	1.06	0.15				0.49
P087679	10	Lower-order	1.08	0.29				0.46
P087684	10	Higher-order	0.77	-0.34				0.58
P087685	10	Higher-order	0.64	0.80				0.40
P087686	10	Higher-order	0.59	3.02				0.17
P103018	8	Lower-order	0.91	0.68			0.37	
P103036	9	Lower-order	0.46	2.50			0.25	
P103049	9	Lower-order	0.89	-0.70			0.63	
P104335	9	Lower-order	1.01	0.47			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 C9: *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.

C.5 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 3.3.1). 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 C12 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 C12: *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
Algebra percent correct	No
Geometry percent correct	Yes
Number sense percent correct	No
Statistics/Reasoning percent correct	Yes
Biology percent correct	No
Chemistry percent correct	No
Physics percent correct	No
Mathematics, below level	No
Mathematics, at level	No
Science, below level	No
Science, at level	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	Yes
Mastery of physics, 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	Yes
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	Yes
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.

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.

Andrabi, T., Das, J., Khwaja, A. I., Vishwanath, T. and Zajonc, T. (2008). *Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to inform the education policy debate*. Research Report 43750, The World Bank, Washington, D.C.

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.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y. and Schady, N. (2016). Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics*, **131** (3), 1415–1453.

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.

ASER (2010). *Annual Status of Education Report 2009 (Rural)*. Tech. rep., Pratham, New Delhi.

ASER (2017). *Annual Status of Education Report (Rural) 2016*. Provisional Report, Pratham, New Delhi.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics*, **60** (1), 47–57.

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

Aucejo, E. M., Romano, T. F. and Taylor, E. S. (2019). *Does Evaluation Distort Teacher Effort and Decisions? Quasi-experimental Evidence from a Policy of Retesting Students*. Working Paper 1612, Centre for Economic Performance, LSE, London.

Azam, M. and Kingdon, G. G. (2015). Assessing Teacher Quality in India. *Journal of Development Economics*, **117**, 74–83.

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., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M. and Walton, M. (2017a). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives*, **31** (4), 73–102.

—, Chassang, S. and Snowberg, E. (2017b). 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., Banerji, R., Duflo, E., Glennerster, R. and Khemani, S. (2010). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. *American Economic Journal: Economic Policy*, **2** (1-30).

—, 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.

Bau, N. and Das, J. (2017). *The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers*. Working Paper 8050, The World Bank, Washington, D.C.

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.

Bergman, P. and Hill, M. J. (2018). The effects of making performance information public: Regression discontinuity evidence from Los Angeles teachers. *Economics of Education Review*, **66**, 104–113.

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

Berry, J., Kannan, H., Mukherji, S. and Shotland, M. (2020). Failure of Frequent Assessment: An Evaluation of India's Continuous and Comprehensive Evaluation Program. *Journal of Development Economics*, **143**, 102406.

Bietenbeck, J., Piopiunik, M. and Wiederhold, S. (2018). Africa's Skill Tragedy Does Teachers' Lack of Knowledge Lead to Low Student Performance? *Journal of Human Resources*, **53** (3), 553–578.

Birdsall, N., Bruns, B. and Madan, J. (2016). *Learning Data for Better Policy: A Global Agenda*. Policy Paper 096, Center for Global Development, Washington, D.C.

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.

Bloom, B. S., Krathwohl, D. R. and Masia, B. S. (1984). *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. New York: Longman, oCLC: 929425977.

Bold, T., Filmer, D., Martin, G., Molina, E., Stacy, B., Rockmore, C., Svensson, J. and Wane, W. (2017). Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa. *Journal of Economic Perspectives*, **31** (4), 185–204.

Bradshaw, L., Izsák, A., Templin, J. and Jacobson, E. (2014). Diagnosing Teachers' Understandings of Rational Numbers: Building a Multidimensional Test Within the Diagnostic Classification Framework. *Educational Measurement: Issues and Practice*, **33** (1), 2–14.

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L. and Yin, Y. (2012). Meta-Analytic Methodology and Inferences About the Efficacy of Formative Assessment. *Educational Measurement: Issues and Practice*, **31** (4), 13–17.

Brookings (2016). *A Global Compact on Learning*. Policy Guide, Brookings Institution, 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.

Buhl-Wiggers, J., Kerwin, J., Smith, J. and Thornton, R. (2017). The Impact of Teacher Effectiveness on Student Learning in Africa. In *Centre for the Study of African Economies Conference*, Munich: cesifo.

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

Burgess, S., Rawal, S. and Taylor, Eric S. (2019). *Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools*. Working Paper 19-139, Annenberg Institute, Providence, RI.

Carril, A. (2017). Dealing with Misfits in Random Treatment Assignment. *The Stata Journal*, **17** (3), 652–667.

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

—, — and — (2010b). *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.

Centro de Estudios (2016). *Base de Datos*.

Chen, J. and de la Torre, J. (2014). A Procedure for Diagnostically Modeling Extant Large-Scale Assessment Data: The Case of the Programme for International Student Assessment in Reading. *Psychology*, **05** (18), 1967–1978.

—, — and Zhang, Z. (2013). Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling: Relative and Absolute Fit Evaluation in CDM. *Journal of Educational Measurement*, **50** (2), 123–140.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, **104** (9), 2633–2679.

Cilliers, J., Fleisch, B., Prinsloo, C. and Taylor, S. (2019). How to Improve Teaching Practice? Experimental Comparison of Centralized Training and In-classroom Coaching. *Journal of Human Resources*, pp. 0618–9538R1.

Clements, D. H. and Sarama, J. (2014). *Learning and Teaching Early Math: The Learning Trajectories Approach*. Studies in mathematical thinking and learning series, New York: Routledge, Taylor & Francis Group, second edition edn.

Confrey, J. (1990). A Review of the Research on Student Conceptions in Mathematics, Science, and Programming. *Review of Research in Education*, **16**, 3–56.

Cortés, F. and Lagos, M. J. (2011). Consecuencias de la Evaluación Docente. In J. Manzi, R. González and Y. Sun (eds.), *La evaluación docente en Chile*, Santiago de Chile: MIDE UC, Centro de Medición Pontificia Universidad Católica de Chile, pp. 137–156.

Cui, Y., Gierl, M. J. and Chang, H.-H. (2012). Estimating Classification Consistency and Accuracy for Cognitive Diagnostic Assessment: Classification Consistency and Accuracy for CDA. *Journal of Educational Measurement*, **49** (1), 19–38.

Dang, H.-A. H. and Glewwe, P. W. (2018). Well Begun, but Aiming Higher: A Review of Vietnam's Education Trends in the past 20 Years and Emerging Challenges. *The Journal of Development Studies*, **54** (7), 1171–1195.

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.

— and Zajonc, T. (2010). India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement. *Journal of Development Economics*, **92** (2), 175–187.

de la Torre, J. (2008). An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications. *Journal of Educational Measurement*, **45** (4), 343–362.

— (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.

— and Chiu, C.-Y. (2016). A General Method of Empirical Q-matrix Validation. *Psychometrika*, **81** (2), 253–273.

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.

de la Torre, 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.

Duflo, A. and Kiessel, J. (2014). *Every child can, every child counts: An evaluation of the Teacher Community Assistant Initiative (TCAI) pilot programme in Ghana*. Working Paper, International Growth Center, London.

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

Dundar, H., Béteille, T., Riboud, M. and Deolalikar, A. (2014). *Student Learning in South Asia: Challenges, Opportunities, and Policy Priorities*. The World Bank.

Educación 2020 (2013). Opinión de Educación 2020 sobre la Evaluación Docente 2012.

Educational Initiatives (2011). *Student Learning Study - An India Report*. Tech. rep., Educational Initiatives, Ahmedabad.

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.

EPDC (2015). *Mapping national assessments*. Policy Brief, FHI 360 Education Policy and Data Center, Washington, D.C.

Escueta, M., Quan, V., Nickow, A. J. and Oreopoulos, P. (2017). *Education Technology: An Evidence-Based Review*. Working Paper 23744, National Bureau of Economic Research.

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?

Filmer, D., Hasan, A. and Pritchett, L. (2006). *A Millennium Learning Goal: Measuring Real Progress in Education*. SSRN Scholarly Paper ID 982968, Social Science Research Network, Rochester, NY.

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.

Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M. and Manzeske, D. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals*. Report NCEE 2018-4001, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

George, A. C. and Robitzsch, A. (2015). Cognitive Diagnosis Models in R: A didactic. *The Quantitative Methods for Psychology*, **11** (3), 189–205.

Gerber, A. S., Arceneaux, K., Boudreau, C., Dowling, C. and Hillygus, D. S. (2015). Reporting Balance Tables, Response Rates and Manipulation Checks in Experimental Research: A Reply from the Committee that Prepared the Reporting Guidelines. *Journal of Experimental Political Science*, **2** (02), 216–229.

Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S. and Santoro, L. E. (2010). Teacher Study Group Impact of the Professional Development Model on Reading Instruction and Student Outcomes in First Grade Classrooms. *American Educational Research Journal*, **47** (3), 694–739.

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.

Graf, E. A. (2008). Approaches to the Design of Diagnostic Item Models. *ETS Research Report Series*, **2008** (1), i–25.

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.

Grissom, J. A. and Youngs, P. (eds.) (2016). *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures*. New York, NY: Teachers College Press.

Hanushek, E. A. and Woessmann, L. (2015). Teach the World. *Foreign Affairs*.

Henson, R. A., Templin, J. L. and Willse, J. T. (2009). Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. *Psychometrika*, **74** (2), 191–210.

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.

Hölmstrom, B. (1979). Moral Hazard and Observability. *The Bell Journal of Economics*, **10** (1), 74–91.

IIPS (2007). *India National Family Health Survey (NFHS-3), 2005-06*. Tech. rep., International Institute for Population Sciences, Mumbai.

Jackson, C. and Cowan, J. (2018). *Assessing the Evidence on Teacher Evaluation Reforms*. Policy Brief 13-1218-1, National Center for Analysis of Longitudinal Data in Education Research, Washington, D.C.

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy*.

—, 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. (2005). Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools. *Journal of Public Economics*, **89** (5), 761–796.

— 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.

Johnson, S. M. and Fiarman, S. E. (2012). The Potential of Peer Review. *Educational Leadership*, **70** (3), 20–25.

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.

Jurich, D. P. and Bradshaw, L. P. (2014). An Illustration of Diagnostic Classification Modeling in Student Learning Outcomes Assessment. *International Journal of Testing*, **14** (1), 49–72.

Kaffenberger, M. and Pritchett, L. (2017). *More School or More Learning? Evidence from Learning Profiles from the Financial Inclusion Insights Data*. Working Paper RISE-WP-17/012, Research on Improving Systems of Education (RISE), Oxford.

Kane, T. J., Blazar, D., Gehlbach, H., Greenberg, M., Quinn, D. and Thal, D. (2019). Can Video Technology Improve Teacher Evaluations? An Experimental Study. *Education Finance and Policy*, pp. 1–55.

Kerwin, J. T. and Thornton, R. L. (2020). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. *The Review of Economics and Statistics*, (Forthcoming), 1–45.

Kingston, N. and Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice*, **30** (4), 28–37.

Koedel, C., Li, J., Springer, M. G. and Tan, L. (2017). The Impact of Performance Ratings on Job Satisfaction for Public School Teachers. *American Educational Research Journal*, **54** (2), 241–278.

—, —, — and — (2019). Teacher Performance Ratings and Professional Improvement. *Journal of Research on Educational Effectiveness*, **12** (1), 90–115.

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

Kraft, M. and Hill, H. (2019). *Developing Ambitious Mathematics Instruction Through Web-Based Coaching: A Randomized Field Trial*. Working Paper 19-119, Annenberg Institute at Brown University, Providence, RI.

Kraft, M. A., Blazar, D. and Hogan, D. (2018a). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*, p. 0034654318759268.

—, Papay, J. P. and Chi, O. L. (2018b). *Teacher skill development: Evidence from performance ratings by principals*. Working Paper, Brown University, Providence, RI.

Kunina-Habenicht, O., Rupp, A. A. and Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, **35** (2-3), 64–70.

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.

—, Zhang, L., Hu, X., Qu, Q., Shi, Y., Qiao, Y., Boswell, M. and Rozelle, S. (2013). Computer assisted learning as extracurricular tutor? Evidence from a randomised experiment in rural boarding schools in Shaanxi. *Journal of Development Effectiveness*, **52** (2), 208–231.

—, —, Qu, Q., Hu, X., Shi, Y., Boswell, M. and Rozelle, S. (2016). *Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in public schools in rural minority areas in Qinghai, China*. Working Paper 237, Stanford University, Freeman Spogli Institute for International Studies, Stanford, CA.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, **76** (3), 1071–1102.

Lee, Y.-W. and Sawaki, Y. (2009). Application of Three Cognitive Diagnosis Models to ESL Reading and Listening Assessments. *Language Assessment Quarterly*, **6** (3), 239–263.

Li, H., Hunter, C. V. and Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, **33** (3), 391–409.

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.

Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L. and Chang, H.-H. (2013). The Development of Computerized Adaptive Testing with Cognitive Diagnosis for an English Achievement Test in China. *Journal of Classification*, **30** (2), 152–172.

—, —, —, — and — (2014). Large-scale implementation of computerized adaptive testing with cognitive diagnosis in China. In Y. Cheng and H.-H. Chang (eds.), *Advancing Methodologies to Support Both Summative and Formative Assessments*, Chinese American Educational Research and Development Association Book Series, Charlotte, NC: Information Age Publishing, pp. 245–262.

Lombardi, M. (2019). Is the remedy worse than the disease? The impact of teacher remediation on teacher and student performance in Chile. *Economics of Education Review*, **73**, 101928.

Louis, K. S. and Marks, H. M. (1998). Does Professional Community Affect the Classroom? Teachers' Work and Student Experiences in Restructuring Schools. *American journal of education*, pp. 532–575.

Lovison, V. and Taylor, E. S. (2018). *Can Teacher Evaluation Programs Improve Teaching?* Technical Report, Stanford University, Stanford, CA.

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, W., Iaconangelo, C. and de la Torre, J. (2016). Model Similarity, Model Selection, and Attribute Classification. *Applied Psychological Measurement*, **40** (3), 200–217.

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

Manzi, J., González, R. and Sun, Y. (2011). *La Evaluación Docente en Chile*. Santiago de Chile: MIDE UC, Centro de Medición Pontificia Universidad Católica de Chile.

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.

McCrary, J. (2008). Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. *Journal of Econometrics*, **142** (2), 698–714.

McMillan, J. H., Venable, J. C. and Varier, D. (2013). Studies of the Effect of Formative Assessment on Student Achievement: So Much More is Needed. *Practical Assessment, Research & Evaluation*, **18**.

Mehta, A. C. (2017). *Elementary Education in India. Analytical Tables 2015-16*. Analytical Tables, District Information System for Education, New Delhi.

Mehta, J. (2013). *The Allure of Order: High Hopes, Dashed Expectations, and the Troubled Quest to Remake American Schooling*. New York: Oxford University Press.

Metcalfe, J. (2017). Learning from Errors. *Annual Review of Psychology*, **68** (1), 465–489.

MHRD (2016). *National Achievement Survey (NAS)*.

Milgrom, P. R. and Roberts, J. (1992). *Economics, Organization, and Management*. New York: Prentice-Hall.

Mo, D., Swinnen, J., Zhang, L., Yi, H., Qu, Q., Boswell, M. and Rozelle, S. (2013). Can one-to-one computing narrow the digital divide and the educational gap in China? The case of Beijing migrant schools. *World Development*, **46**, 14–29.

MoEYS (2015). *Results of Grade Six Student Achievement from the National Assessment in 2013*. Report, Ministry of Education, Youth and Sport Cambodia, MoEYSEducation Quality Assurance Department, EQAD, Phnom Penh.

Molina, E., Fatima, S. F., Ho, A., Hurtado, C. M., Wilichowksi, T. and Pushparatnam, A. (2018). *Measuring Teaching Practices at Scale: Results from the Development and Validation of the Teach Classroom Observation Tool*. Working Paper 8653, The World Bank, Washington, D.C.

MoPME (2014). *National Student Assessment 2013 for Grades 3 and 5*. Tech. rep., Monitoring & Evaluation Division, Directorate of Primary Education Ministry of Primary & Mass Education, Dhaka.

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.

— and Zieleniak, Y. (2013). Chasing the Syllabus: Measuring Learning Trajectories in Developing Countries with Longitudinal Data and Item Response Theory.

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.

National Education Association (2019). *Teacher Assessment and Evaluation: The National Education Association's Framework for Transforming Education Systems to Support Effective Teaching and Improve Student Learning*. White Paper, National Education Association, Washington, D.C.

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

NEREC (2009). *National Assessment of Achievement of Students Completing Grade 4 in Year 2009*. World Bank and Ministry of Education Assisted Publication, University of Colombo, Colombo.

Papay, J. (2012). Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation. *Harvard Educational Review*, **82** (1), 123–141.

Papay, J. P., Taylor, E. S., Tyler, J. H. and Laski, M. (2019). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*.

Pearlman, L. (2011). *Common and Textbook Foil Groupings: A Social Network Approach to Distractor Analysis*. ProQuest LLC.

Pope, N. G. (2019). The effect of teacher ratings on teacher performance. *Journal of Public Economics*, **172**, 84–110.

Pritchett, L. and Beatty, A. (2015). Slow Down, You're Going Too Fast: Matching Curricula to Student Skill Levels. *International Journal of Educational Development*, **40**, 276–288.

Proctor, C. H. (1970). A Probabilistic Formulation and Statistical Analysis of Guttman Scaling. *Psychometrika*, **35** (1), 73–78.

Ramachandran, V., Béteille, T., Linden, T., Dey, S., Goyal, S. and Goel Chatterjee, P. (2018). *Getting the Right Teachers into the Right Schools: Managing India's Teacher Workforce*. Washington, D.C.: The World Bank.

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

Rolleston, C. (2014). Learning Profiles and the 'Skills Gap' in Four Developing Countries: A Comparative Analysis of Schooling and Skills Development. *Oxford Review of Education*, **40** (1), 132–150.

Rupp, A. A., Templin, J. and Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications*. Methodology in the social sciences, New York: Guilford Press, oCLC: ocn424561376.

— and Templin, J. L. (2008). Unique Characteristics of Diagnostic Classification Models: A Comprehensive Review of the Current State-of-the-Art. *Measurement: Interdisciplinary Research & Perspective*, **6** (4), 219–262.

Saavedra, J., Näslund-Hadley, E. and Alfonso, M. (2017). *Targeted Remedial Education: Experimental Evidence from Peru*. Tech. Rep. w23050, National Bureau of Economic Research, Cambridge, MA.

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.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in science Teaching*, **35** (3), 265–296.

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

Sandefur, J. (2018). Internationally Comparable Mathematics Scores for Fourteen African Countries. *Economics of Education Review*, **62**, 267–286.

Santiago, P., Benavides, F., Danielson, C., Goe, L. and Nusche, D. (2013). *Teacher Evaluation in Chile*. Paris: Organisation for Economic Co-operation and Development.

Sartain, L. and Steinberg, M. P. (2016). Teachers' Labor Market Responses to Performance Evaluation Reform: Experimental Evidence from Chicago Public Schools. *Journal of Human Resources*, **51** (3), 615–655.

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.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, NY: Oxford University Press, google-Books-ID: eDWG3728OxcC.

Singh, A. (2014). *Emergence and Evolution of Learning Gaps across Countries. Panel Evidence from Ethiopia, India, Peru and Vietnam*. Working Paper 124, University of Oxford, Oxford.

— (2019). Learning More with Every Year: School Year Productivity and International Learning Divergence. *Journal of the European Economic Association*, pp. 1–44.

Sorrel, M. A., de la Torre, J., Abad, F. J. and Olea, J. (2017). Two-Step Likelihood Ratio Test for Item-Level Model Comparison in Cognitive Diagnosis Models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **13** (Supplement 1), 39–47.

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.

Steinberg, M. P. and Sartain, L. (2015). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, **10** (4), 535–572.

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

Tatsuoka, K. K., Corter, J. E. and Tatsuoka, C. (2004). Patterns of Diagnosed Mathematical Content and Process Skills in TIMSS-R Across a Sample of 20 Countries. *American Educational Research Journal*, **41** (4), 901–926.

Taut, S., Santelices, M. V., Araya, C. and Manzi, J. (2011). Perceived Effects and Uses of the National Teacher Evaluation System in Chilean Elementary Schools. *Studies in Educational Evaluation*, **37** (4), 218–229.

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

— and Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *The American Economic Review*, **102** (7), 3628–3651.

The World Bank (2017a). *Learning to Realize Education's Promise. World Development Report 2018*. Washington, D.C.: The World Bank, oCLC: 992735784.

The World Bank (2017b). *World Development Report Concept Note*. Concept Note, The World Bank, Washington, D.C.

The World Bank (2018a). *India: Systematic Country Diagnostic*. Country Diagnostic 126284-IN, The World Bank, Washington, D.C.

The World Bank (2018b). *World Development Indicators. Data*.

Tu, D., Gao, X., Wang, D. and Cai, Y. (2017). A New Measurement of Internet Addiction Using Diagnostic Classification Models. *Frontiers in Psychology*, **8**.

UNESCO (2015). *EFA Global Monitoring Report, 2015. Education for All 2000-2015: Achievements and Challenges*. Paris, France: UNESCO, 2nd edn.

UNESCO Institute for Statistics (2018). *Data for the Sustainable Development Goals*.

USAID (2017). Early Grade Reading Barometer.

VanLehn, K. (1990). *Mind bugs: the origins of procedural misconceptions*. Learning, development, and conceptual change, Cambridge, Mass: MIT Press.

Vivalt, E. (2017). *How Much Can We Generalize From Impact Evaluations?* Job market paper, Australian National University, Canberra, Australia.

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.

Wonu, N. and Zalmon, I. G. (2017). Diagnosis and Remediation Of Senior Secondary Students' Common Learning Difficulties In Mathematics From Chief Examiners' Report. *European Journal of Research and Reflection in Educational Sciences*, **5** (1), 7–23.

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.