The Relative Effectiveness of Teachers and Learning Software: Evidence from a Field Experiment in El Salvador

Konstantin Büchel[a], Martina Jakob[b],
Christoph Kühnhanss[b], Daniel Steffen[c], Aymo Brunetti[a]

[a]Department of Economics, University of Bern
[b]Institute of Sociology, University of Bern
[c]Institute of Financial Services, Lucerne University of Applied Sciences and Arts

August 30, 2021

Abstract

This study provides evidence on the relative effectiveness of computer-assisted learning (CAL) software and traditional teaching. Based on a field experiment in Salvadoran primary schools, we evaluate three interventions that aim to improve learning in mathematics: *(i)* additional teacher-led classes, *(ii)* additional CAL classes monitored by a supervisor, and *(iii)* additional CAL classes instructed by a teacher. We find that CAL lessons lead to larger learning gains and are less sensitive to class size as well as student ability than teacher-centered classes. Our results highlight the value of CAL in an environment with heterogeneous classes and poorly qualified teachers.

We are grateful to David Burgherr, Malin Frey, and Amélie Speiser who provided excellent research assistance and to Philippe Sasdi for coordinating data collection in Swiss primary schools. The project further benefited from invaluable feedback by Michael Gerfin, Ben Jann, Florian Keller, Ulf Liebe, Blaise Melly, Urs Moser, Adina Rom, Mauricio Romero, Erik Snowberg, and the participants at the NADEL Workshop (ETHZ), Brown Bag Seminar (Department of Economics, U. Bern), CRED Seminar (U. Bern), SEVAL Meeting (hosted by the SDC in Bern), Rational Choice Sociology Conference in Venice, the SSES Annual Congress in Geneva, the EEA Annual Meeting 2020, and the Annual DENS Meeting in St. Gallen. This study would not have been possible without the *Impact Award 2017* prize money awarded by the SDC and NADEL (ETHZ), and funding by the chair of Ben Jann (Institute of Sociology, University of Bern) as well as the IMG Stiftung. Martina Jakob and Christoph Kühnhanss disclose that they serve on a voluntary basis as president and vice-president of *Consciente – Unterstützungsverein für El Salvador (Schweiz)*. We received IRB approval from the Faculty of Business, Economics and Social Sciences at the University of Bern. A randomized controlled trials registry entry is available at: `https://www.socialscienceregistry.org/trials/2789`.

Contact Details (°Corresponding Author):
°Büchel: Univ. of Bern, Dept. of Economics, Schanzeneckstr. 1, CH-3001 Bern, konstantin.buechel@vwi.unibe.ch
Jakob: Univ. of Bern, Inst. of Sociology, Fabrikstr. 8, CH-3012 Bern, martina.jakob@soz.unibe.ch
Kühnhanss: Univ. of Bern, Inst. of Sociology, Fabrikstr. 8, CH-3012 Bern, christoph.kuehnhanss@soz.unibe.ch
Steffen: Lucerne University of Applied Sciences and Arts, Suurstoffi 1, CH-6343 Rotkreuz, dani.steffen@hslu.ch
Brunetti: Univ. of Bern, Dept. of Economics, Schanzeneckstr. 1, CH-3001 Bern, aymo.brunetti@vwi.unibe.ch

1 Introduction

While net primary school enrollment rates in low-income countries climbed from 56% in 2000 to 81% in 2019, learning outcomes have failed to keep pace. Less than 15% of primary school children in low-income countries pass minimum proficiency thresholds in reading and math, compared to about 95% of pupils in high-income countries (World Bank, 2018, p. 8). Public schooling systems in developing countries face multiple challenges that curb their productivity. These include a mismatch between national curricula and student abilities (Pritchett and Beatty, 2015), large and heterogeneous classes (Mbiti, 2016; Glewwe and Muralidharan, 2016), and low levels of effort among poorly qualified teachers (Chaudhury et al., 2006; Bold et al., 2017a). A much-noticed approach to overcome these barriers is the use computer-assisted learning software (e.g. The Economist, 2017). Computer-assisted learning (CAL) has several advantages over traditional teaching methods, as it allows for self-paced learning that is tailored to the abilities of the student, provides instant feedback and is less sensitive to the motivation and skills of teachers. Previous studies on the impact of technology-based teaching methods on learning outcomes are encouraging. CAL interventions are usually found to improve students' test scores and seem to be particularly beneficial to advance learning in math (for a comprehensive review see Escueta et al., 2020).[1]

Yet, most studies evaluate CAL lessons that were offered as a supplement to regular classes, meaning that beneficiaries experienced a considerable expansion of school time compared to the untreated students in the control group. Thus, it remains unclear whether learning gains are

[1] Experimental studies on CAL interventions in low- and middle-income countries include Banerjee et al. (2007), Linden (2008), Carrillo et al. (2010), Yang et al. (2013), Mo et al. (2015), Lai et al. (2015), Muralidharan et al. (2019), and Bettinger et al. (2020). All eight studies report positive intent-to-treat estimates on learning outcomes in the range of 0.1σ to 0.4σ and are discussed in appendix D. Comparing these results to the evaluation of the *One Laptop per Child Initiative* in Peru (Beuermann et al., 2015; Cristia et al., 2017) suggests that CAL software is an essential ingredient to computer-based learning whereas "bare" hardware provides no educational value.

actually attributable to the use of the software or if additional lessons conducted by a teacher might have produced similar or even better results.[2] In addition, there is little evidence on whether CAL can be seen as a substitute for teachers or if it is a complement to them. Finally, previous research has mostly evaluated specifically customized software which is only available in a limited number of languages. As a result, many policy-makers with an interest in implementing CAL cannot draw on software that is readily available and has been successfully evaluated.

Based on a randomized controlled trial, this paper examines the relative effectiveness of primary school math teachers and a freely available CAL software that features content in more than 30 languages. To disentangle the effects of additional teaching and the use of a learning software, the experimental design features three treatments that did not interfere with regular lessons: The first treatment comprises additional, i.e. *after* class, math lessons instructed by a contract teacher (henceforth labeled as TEACHER). The second and third treatments are additional, i.e. *after* class, math lessons based on CAL software; one group of CAL classes is monitored by technical supervisors (CAL+SUPERVISOR), while the other group is taught by contract teachers (CAL+TEACHER). Contract teachers had to be officially certified to teach math in primary schools, whereas supervisors were laypersons instructed to provide no content-related help. In CAL lessons, each child worked

[2] To our knowledge, the only study that evaluates the effectiveness of CAL lessons as a substitute to teacher-centered lessons in the development context was conducted by Linden (2008) in India. While attending *additional* CAL lessons raised math scores of second and third graders, CAL had a negative impact when it *substituted* regular classes. As the author points out, the study examines a sample of NGO-run schools with well trained staff and innovative teaching methods. At the same time, the evaluated software did neither personalize content nor were the CAL sessions geared towards self-paced learning. While it is unclear whether these findings translate to the use of modern software or the challenging contexts of public education systems in developing countries, they still raise doubts about the inherent benefits of technology-based instruction. A more promising result is documented by Bettinger et al. (2020) on the value of CAL-based homework assignments. For a sample of Russian primary schools, they show that at low dosage levels, CAL homework yields larger learning gains than traditional homework tasks. For a more comprehensive review, we refer to appendix D.

individually on a laptop using KHAN ACADEMY content via the offline application KA LITE. In teacher-centered lessons, children were arranged in table groups where they worked on content tailored to their ability. Each of the three treatments comprised two 90 minutes math lessons per week over a six months period and were implemented by the Swiss-Salvadoran NGO CONSCIENTE.

We conducted the field experiment between February and October 2018 with a sample of 198 primary school classes spanning grades three to six in the rural district of Morazán, El Salvador. Twenty-nine out of 57 eligible schools were randomly selected for program participation. The 158 classes from these 29 schools were then randomly assigned to either Treatment 1 (i.e. TEACHER, 40 classes), Treatment 2 (i.e. CAL+SUPERVISOR, 39 classes), Treatment 3 (i.e. CAL+TEACHER, 39 classes) or a program school control group (i.e. CX: CONTROL S.T. EXTERNALITIES, 40 classes). In the 28 non-program schools, a random sample of 40 classes was drawn resulting in a "pure" control group that is not subject to potential treatment externalities.

Our analysis establishes four major findings. *First*, the additional CAL classes had a considerable impact on students' math skills. Being assigned to additional CAL lessons increased their math scores by 0.21σ (p-value<0.01) when overseen by a supervisor and by 0.24σ (p-value<0.01) when instructed by teachers. These intent-to-treat estimates, which reflect a program attendance rate of 59%, correspond to the average increase in math abilities over 0.6 school years. Using the treatment assignment as instrumental variable for attendance, we estimate that participating in all 46 additional CAL lessons translates to average learning gains of 0.38σ and 0.40σ, respectively. This is equivalent to the average increase in math abilities during 1.1 school years.

Second, additional CAL lessons were more productive than the additional math lessons instructed by a teacher. The intent-to-treat effect of being assigned to additional teacher-led classes without CAL was 0.15σ (p-value=0.02). Hence, students assigned to CAL+TEACHER outperformed students assigned to TEACHER by 0.09σ (p-value=0.10). The CAL treatment overseen by technical supervisors (CAL+SUPERVISOR) was also more successful in raising student learning than traditional teaching, even though this difference is not statistically significant (p-value=0.24). The advantage of CAL lessons relative to teacher-centered lessons was most pronounced in the domain of number sense and elementary arithmetic, and less so with respect to geometry, measurement and data. Focusing on number sense and elementary arithmetic, the difference between the CAL and non-CAL treatments increases to 0.11σ (p-value=0.06) for CAL instructed by teachers and to 0.09σ (p-value=0.12) for the CAL monitored by supervisors.

Third, we present nuanced evidence on the specific strengths of CAL-based learning. Math instruction via CAL software did not only marginally outperform teacher-centered lessons but its effect was also less sensitive to class size and the initial ability of participants. For instance, evaluating the impact of each treatment for students attending classes with 27 students (i.e. equivalent to a class size of one standard deviation above the mean), we obtain effects of -0.01 (p-value=0.87) for TEACHER, 0.17 (p-value=0.01) for CAL+SUPERVISOR, and 0.14 (p-value=0.04) for CAL+TEACHER. Against the background of typically large and heterogeneous classes in developing countries, the relatively homogeneous impact of CAL-based lessons along class size and student ability underscores the educational value of CAL approaches.

Compared to the two CAL interventions, the impact of traditional teaching was also more sensitive to the grade level of the assigned classes. In fact, the observed gap in learning gains between teacher-centered lessons and CAL instructions primarily accrued among students of grade five and six. To better understand this pattern, we discuss a teacher math assessment covering the primary school curriculum of El Salvador. Results from this assessment show that the performance of the employed contract teachers drops from an average of 75% correct answer on items from grade three to 54% correct answers on items from grade six. Regular math teachers in local primary schools are even less proficient in their subject than the contract teachers hired by the NGO. While the median contract teacher correctly answered 66% of 50 questions spanning grade two to grade six, the median regular teacher correctly solved 44% of the same items. Potential productivity gains resulting from an introduction of CAL to regular classes would, therefore, likely exceed our estimates based on the comparison between the CAL treatments and the TEACHER treatment.

Fourth, we document substantial treatment externalities. At endline, students in program school control classes outperformed students in pure control classes by 0.14σ (p-value=0.02), although they were only indirectly exposed to the three treatments. In particular, we find evidence for spillovers from the two CAL treatments. While we cannot comprehensively pin down the mechanisms at work, suggestive evidence points toward peer effects. At the same time, the data rejects hypotheses operating via direct exposure of students in control classes to the treatments (i.e. non-compliance) or testable hypotheses on behavioral adjustments in response to the experimental design.

This study makes several contributions to the literature on educational interventions in developing countries. *First*, it improves our understanding of how CAL performs relative to alternative teaching models. As opposed to Linden (2008), who documents a negative value-added of CAL in NGO-administered schools in India, our findings suggest that CAL has the potential to outperform traditional teacher-led instruction, especially if teachers are poorly qualified. According to our estimates, it would take a teacher at the 91th percentile of the local teacher ability distribution to achieve the same learning gains as observed in CAL lessons overseen by a supervisor. This corresponds to a teacher in the 75th ability percentile among the contract teachers hired for this experiment or to 88% correct answers in the administered teacher assessment. While CAL has been praised in terms of its individualized pedagogy (e.g. Banerjee et al., 2007; Muralidharan et al., 2019), these numbers highlight that it may also be a promising approach to mitigate the adverse effects of teachers' inadequate content and pedagogical knowledge, as it has been recently documented for several developing countries (e.g. Bold et al., 2017a).

Second, we present the first experimental test of complementarities between teachers and learning software. In our setting, teachers play a marginal role in the success of technology-based instruction, with CAL lessons being almost equally effective when conducted by a supervisor instead of an officially certified teacher. While the interplay between software and teachers likely depends on many factors, such as the distribution of teacher ability, the teachers' mandate, and the mix of concepts taught in class, this field experiment yields no indication for substantive complementarities. The analysis rather suggests that learning software can be a promising substitute to regular teaching in an environment with poorly qualified teachers. Only few experimental studies aspire to distinguish between complementary and substitutable inputs entering the educational production function; notable exceptions are recent papers by Mbiti et al. (2019) on the complementarity between school grants and teacher incentives in Tanzanian primary schools, and by Attanasio et al. (2014) on the complementarity between psychosocial stimulation programs and nutritional supplements in early childhood development.

Third, we contribute to the broader literature on treatment externalities (e.g. Miguel and Kremer, 2004; Baird et al., 2018). By including control classes from treatment schools as well as spatially separated pure control classes from non-treatment schools into our experimental design, this study provides a reasonable identification of potential externalities. Our findings underscore the importance of appreciating the possibility of externalities in the design of experimental evaluation studies, even when such effects appear unlikely at first sight. Moreover, the presence of positive treatment externalities provide a strong rationale in favor of scaling the evaluated program.

Finally, this study adds to the accumulated evidence on the effectiveness of CAL by evaluating a widely available off-the-shelf software. In contrast to software tested in previous evaluations, Khan Academy is free of charge and features extensive math content in more than 30 languages.[3]

[3] The full version is available in 16 languages including Spanish, and a subset of content is available in about 20

Like in many developing countries, poor internet coverage is a challenge in El Salvador. We therefore deployed an open-source platform, KA Lite, specifically designed to make offline learning with Khan Academy content possible. Using KA Lite's excellent monitoring features, we maintained a semi-autonomous feedback-loop during CAL lessons allowing for individually tailored and self-paced learning. Our results on the performance of this specific CAL setup, including the homogeneous impact along class size and student ability, shows how a non-adaptive software can be successfully utilized to individualize instructions. Since the deployed software is arguably one of the characterizing features of a CAL intervention, our findings bear direct policy relevance for educational decision-makers around the globe who are looking for a readily available learning software suitable in non-English speaking countries with poor internet coverage.

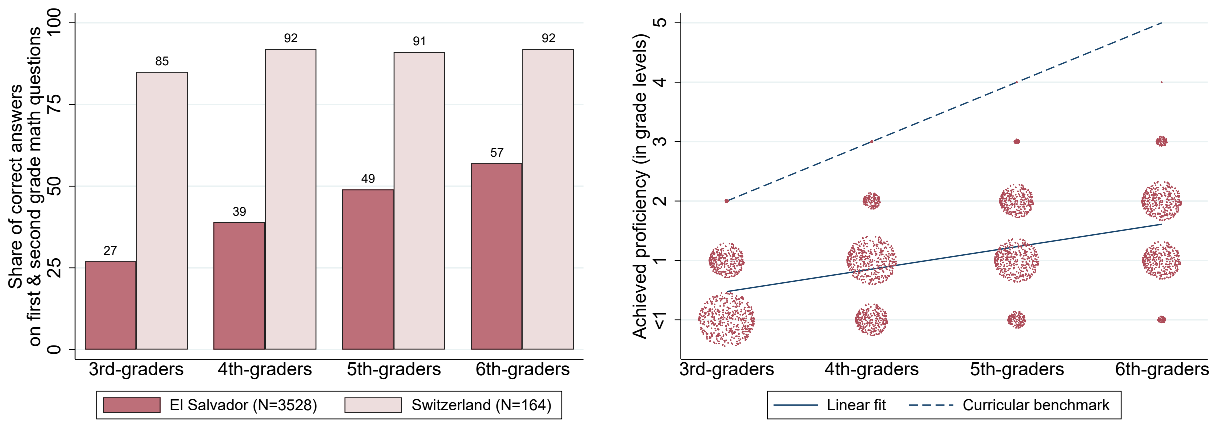
2 Context and Intervention

El Salvador is a lower middle-income country in Central America. The country's net primary enrollment rates are estimated at 80%, which is 7 percentage points below the average of lower middle-income countries. While most children attend primary school, enrollment declines to 67% at the secondary level and to 28% in tertiary education.[4]

The department of Morazán is a poor and rural region in the northeast of the country with roughly 200,000 inhabitants. An average person in Morazán lives on 3.80 USD per day and, according to national definitions, almost 50% of the households face multifaceted poverty. While Morazán ranks second-last among all Salvadoran departments in terms of adult literacy, its secondary school

languages. Another off-the-shelf learning software that has been successfully evaluated is Mindspark (see Muralidharan et al., 2019), which operates in English and Hindi for math and language training.

[4]Enrollment statistics according to the *World Development Indicators* provided online by the World Bank, see `https://data.worldbank.org/indicator` (last access: 26.10.2019)



(a) Share of correct answers on 1st/2nd grade math questions among Salvadoran and Swiss pupils.

(b) Assessed grade level in math among third to sixth graders in Morazán early in their school year.

Figure 1: Math learning outcomes in Morazán (Panels a & b) and Switzerland (Panel a).
Note: Panel (b) illustrates the achieved proficiency in math (measured in grade levels) among third to sixth graders in Morazán at the beginning of their school year. A student, each represented by a dot, needs to score at least 50% correct answers on grade specific items in order to reach the next proficiency level. Since the test was administered in the first weeks of their school year, a third grader answered first and second grade items and therefore may be assigned to grade level 2, 1 or <1 depending on her performance. The size of the bubbles are proportional to the number of students they represent. Further explanations are provided in appendix A.1. *Source:* Baseline data, February 2018.

students came forth in the 2018 "PAES" national examinations (Digestyc, 2018; MINED, 2018).

Our math assessments with 3,528 third to sixth graders conducted in February 2018 reveal that primary school children in Morazán barely grasp the most elementary concepts in math. Figure 1a shows that the share of correct answers to first and second grade questions increases from 27% among third graders to 57% among sixth graders, who by then should have attended more than 1,000 math lessons. To put these numbers into context, we conducted the same test with 164 pupils in Switzerland, who answered on average between 85% and 92% of the items correctly. Even the worst performing Swiss third grader outperformed the median sixth grader in Morazán. Similarly flat learning curves among primary school children have also been reported for other low- and middle-income regions across the globe, including countries in Western Africa, Eastern Africa, Central America, and South East Asia (e.g. Beatty et al., 2018; PAL, 2020).

Several challenges that plague Morazán's schooling system can account for its low productivity. For instance, our monitoring data from school visits reveal high rates of teacher absenteeism suggesting that, on average, 25% of regular lessons are canceled. Low teacher motivation mixes with outdated pedagogical techniques that essentially follow the logic of "copy, memorize, and reproduce". And despite relatively small class sizes – the pupil-teacher ratio averages 28-to-1 at the national level and 19-to-1 in our sample – heterogeneous student performance and an overambitious curriculum make it difficult to teach at an appropriate level. As Figure 1b shows, third to sixth graders lag considerably behind the official curriculum and this gap widens as children move up to higher grade levels. Moreover, performance heterogeneity within classes is considerable. In the majority of classes, students' math ability spans three grades or more (for further explanations see appendix A.1). In general, the public schooling system in El Salvador faces challenges similar to those reported for other low- and middle income countries.[5]

The Salvadoran Ministry of Education has recently put considerable effort into addressing learning deficiencies in public schools. While primary schooling has been typically confined to either morning or afternoon lessons throughout El Salvador, a recent policy aims to extend school time over a full day and to complement traditional teaching with innovative learning approaches (MINED, 2013). The government hopes that longer schooldays will not only boost learning outcomes, but also shield children from the influence of gangs. Within the scope of this countrywide program, the Ministry of Education seeks to cooperate with NGOs in order to collectively promote a flexible

[5] The pupil-teacher ratio in middle-income countries averages 24-to-1, while it climbs to 40-to-1 in low income countries (Unesco, 2019); in some contexts, such as rural India, it can even reach 90-to-1 (Mbiti, 2016). Besides the large class size, students' abilities and preparation levels are often very heterogeneous, which is also the case in our data. For example, Muralidharan et al. (2019) report for their sample of 116 Indian middle schools that students' ability in the median classroom spans four grades in both math and language, while we obtain three grade levels for primary schools. Moreover, Pritchett and Beatty (2015) show that the pace of learning is very slow in developing countries and that there is a mismatch between curriculum and student abilities. This is consistent with what we observe in Figure 1b. Finally, low teacher motivation is a well-known issue: Chaudhury et al. (2006) find that 19% of teachers in developing countries are absent during unannounced visits, while our monitoring data suggests that 25% of classes in Morazán's primary school are canceled.

curriculum. While all schools received instructions to expand their school days, most of them have not put the policy into practice due to a lack of resources to pay for further teaching staff.

Intervention. In this context, we evaluate the impact of an educational initiative on math abilities of primary school children of grades three to six.[6] The program features three intervention arms consisting of two additional math lessons per week that take place either before or after regular classes. Each lesson has a duration of 90 minutes, meaning that the beneficiaries' number of math lessons are almost doubled during the program phase. The first intervention arm comprises additional math lessons instructed by a teacher without using software. The second and third intervention arms are additional math lessons based on CAL software; one group of classes is taught by teachers, while the other group is monitored by supervisors. The additional classes were not mandatory, but participation was officially recommended by the Ministry of Education.

The *CAL-lessons* in the second and third intervention arm were based on an offline application of the learning platform Khan Academy, which is known as KA Lite. This freely available software provides a wide range of instructional videos and exercises for every difficulty level. While the learning tool is not directly adaptive, it allows teachers to track the progress of each student and assign appropriate contents based on prior performance. To tailor instruction to students' learning levels, a set of workplans covering different content units was prepared. Based on a placement test, children received a plan that was viewed as adequate for their respective level and they could then proceed to subsequent plans at their own pace. Since one computer was available per student, each child could follow its individual learning path. Typically, students started with materials from lower grades and then slowly progressed towards contents corresponding to their actual grade level.

A similar methodology was used for the first intervention arm that features more traditional

[6] This paragraph summarizes the characterizing features of the three treatment arms. More comprehensive information on the design of the intervention is provided in the appendix section C.

math lessons instructed by a teacher. According to their initial math skills, children were arranged in table groups where they worked on plans tailored to their ability. Teachers were instructed to explain important concepts, correct students' work at home and promote children (or entire table groups) to subsequent plans when appropriate. While this strategy only allows for a crude approximation of teaching to each child's ability level, it represents a degree of individualization that can realistically be achieved without the help of technology.

To pay credit to the *social component of learning*, all treatments combined individualized learning with educational games. For this purpose, a manual containing animation and math games was developed. The manual compiles simple techniques to promote students' collective learning as well as their motivation to participate in class. Games were usually played at the beginning, in the middle, or at the end of each session. While supervisors were instructed to use animation and concentration games, teachers were additionally introduced to a series of math games.

The contracted *teachers* were required to be officially certified to instruct grades three to six in math. That is, they all possessed a university degree and had either completed a teacher education, or another study program combined with a one-year pedagogical course. Teachers were selected based on a brief math assessment and a job interview. They were employed on short-term contracts and earned 300 USD per month for teaching additional math lessons to four classes.[7] For lessons that were canceled, teachers received no remuneration. To optimize the comparability of treatments, all teachers were assigned an equal number of CAL and non-CAL lessons. Before and during the intervention, teachers were trained to operate the learning software and they reviewed mathematical concepts as well as central pedagogical strategies including the use of educational games. Teaching was tightly monitored by our partner NGO through monthly feedback meetings at the NGO's headquarters and unannounced classroom visits during the implementation phase.

[7] This corresponds to 8×90 minutes of teaching per week, or – including preparatory work – to a 60% job.

The *supervisors* received only technical training and were paid substantially less than teachers, that is 180 USD for taking care of four classes. They were required to have minimal IT skills and some experience in dealing with children, but no contracted supervisor possessed a degree in education or teaching credentials. During the intervention, supervisors were instructed to restrain from providing any content-specific help. Like teachers, supervisors were employed on short-term contracts and were paid conditional on the number of classes they conducted.

3 Research Design

This study is built around an RCT to identify the causal impact of the three interventions arms. It started in February 2018 with a baseline assessment and a survey covering all control and program classes. The additional math classes began in April 2018 and were implemented until the end of the school year in fall 2018; the school year in El Salvador starts in mid-January and ends in November. The endline tests took place in October 2018, six months after the start of the intervention. Again, all program and control classes took part in the endline tests.

3.1 Sampling and Randomization

Our sampling and randomization scheme has three layers, as exemplified in Figure 2. Starting point are all 302 primary schools in Morazán. In coordination with the NGO and the regional Ministry of Education, we defined the following eligibility criteria for a preselection of primary schools:

- ***School size***, *eliminates 221 schools:* A school was considered too small, if it had integrated classes (across grades) or gaps in its grade structure (i.e. not at least one class per grade). This guarantees that every eligible school has at least four different classes in grades three to six, and therefore can participate with at least (*i*) one CAL+TEACHER, (*ii*) one CAL+SUPERVISOR, (*iii*) one TEACHER, and (*iv*) one control class;

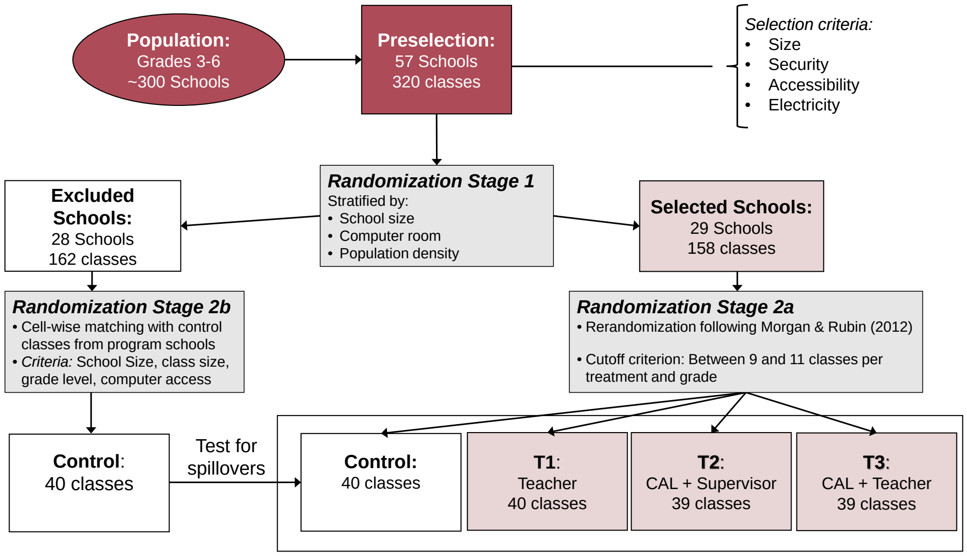


Figure 2: Sampling and randomization scheme.

- ***Security**, eliminates 14 of the remaining 81 schools:* Based on an assessment by the local staff and the regional Ministry of Education, schools located in areas dominated by criminal gangs were excluded due to security concerns;

- ***Accessibility**, eliminates 7 of the remaining 67:* Schools where access by car is difficult were discarded. To inform this decision we relied on Google-Maps driving times and a validation by local staff and the regional Ministry of Education;

- ***Electricity**, eliminates 3 of the remaining 60 schools:* Schools without a (close-by) power supply did not qualify for the program.

After this pre-selection, 57 schools with a total of 320 eligible classes and about 6,400 students remained in the sample. In *randomization stage 1*, 29 of the 57 schools were randomly chosen to participate in the program. To improve balance, the assignment was stratified by school size, local population density and students' access to a computer room.

In *randomization stage 2a*, we randomly assigned the 158 classes in the 29 selected program schools to the control group or one of the three intervention arms. Following Morgan and Rubin

(2012) we re-run the randomization routine until the interventions were balanced across schools and grades. This mechanism assigned 39 classes to CAL+TEACHER, 39 classes to CAL+SUPERVISOR, 40 classes to TEACHER, and 40 classes to the control group.[8] We account for this procedure when comparing estimates within program schools by computing randomization inference test statistics based on 2,000 random draws subject to the identical cut-off criterion. Our choice to run 2,000 draws is guided by Young (2019), who finds no appreciable change in rejection rates beyond this threshold. To implement randomization tests we rely on Stata's RITEST-package developed by Heß (2017).

As prominently discussed in Miguel and Kremer (2004), interventions can have spillover effects on non-participating students from the same school or area. A unique feature of our design allows us to estimate the size of such treatment externalities. For this purpose, in *randomization stage 2b*, 40 additional control classes from non-treatment schools were included in the study. These additional "pure" control classes are spatially separated from the intervention, and thus not affected by the NGO's work. The *pure control classes* were randomly selected from the 28 control schools by matching them cell-wise to the distribution of control classes from program schools, accounting for school size, grade level, class size and students' access to computers.

This procedure yields five different groups of primary school classes, namely the 39 or 40 classes assigned to each of the three treatment groups, 40 control classes from the 29 program schools, and 40 pure control classes from the 28 control schools.

3.2 Data

In the course of the evaluation, four types of data were gathered: (*i*) Math learning outcomes of students were assessed before and after the intervention, (*ii*) socio-demographic statistics stem from

[8]For reasons related to the timing and implementation of the program, we had to run randomization before processing the baseline assessment and were thus unable to stratify by students' test scores.

a survey that children answered prior to the baseline math assessment, (*iii*) administrative data on schools was collected between October 2017 and February 2018, and (*iv*) monitoring data was recorded during unannounced school visits throughout the program phase. Table 1 shows summary statistics for the main variables collected before the start of the program as well as absence rates at the endline and baseline assessment. In particular, it displays means and standard errors for the different variables by treatment status, and tests whether the mean is equal across the five groups.

While the treatment and control groups do not differ significantly in any observable dimension at baseline, Table 1 shows a sizeable increase in the absence rates between the baseline and the endline assessment. Before both rounds of data collection, we updated comprehensive class lists of registered pupils. This revealed that large numbers of children either migrated out of Morazán or discontinued their education. We achieved an attendance of about 95% registered pupils in both rounds, but since classes shrank during the school year, the overall attrition at endline almost hits the 10% mark. Importantly, Table 1 does not point toward systematic differences in attrition rates by treatment status.[9] Moreover, compliance with the experimental protocol was very good in the sense that only 38 out of 3197 students (i.e. 1.2%) within our estimation sample switched between different classes, grades or schools.

3.2.1 Math Learning Outcomes

The math assessments include 60 items covering the primary school curriculum of El Salvador. The weighting of questions across the three main topics (a) number sense & elementary arithmetic (∼65%), (b) geometry & measurement (∼30%), and (c) data & statistics (∼5%) was closely aligned with the national curriculum. Moreover, we verified the appropriateness of each question

[9]We examine this more closely in Table A.1 in the appendix, confirming that the treatment status is not significantly correlated with presence at the endline test.

Table 1: Balance at baseline and absence rates during assessments

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Math Scores (N=3528)	T1: Math w. Teacher	T2: CAL w. Supervisor	T3: CAL w. Teacher	Program School Controls	Pure Control Classes	p-value
%-Share Correct Answers	30.33	33.47	31.97	32.60	30.80	0.45
	(1.80)	(1.90)	(2.07)	(1.32)	(2.00)	
IRT Math Score	0.01	0.18	0.08	0.08	0.00	0.72
	(0.14)	(0.14)	(0.16)	(0.10)	(0.15)	
Panel B: Sociodemographics (N=3528)						
Female Student	0.50	0.52	0.55	0.51	0.49	0.43
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	
Student Age	−0.09	−0.01	0.02	−0.03	−0.03	0.70
	(0.08)	(0.09)	(0.09)	(0.06)	(0.09)	
Household Size	5.56	5.61	5.57	5.55	5.50	0.92
	(0.13)	(0.12)	(0.12)	(0.08)	(0.12)	
Household Assets Index	0.55	0.55	0.54	0.56	0.56	0.88
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	
Panel C: Class Room Variables (N=198)						
Class Size	18.40	19.33	18.69	18.13	18.32	0.92
	(1.37)	(1.35)	(1.37)	(0.96)	(1.54)	
Female Teacher	0.80	0.77	0.77	0.72	0.55	0.14
	(0.10)	(0.10)	(0.10)	(0.07)	(0.11)	
Absence Rate at Baseline (%)	3.88	3.15	5.39	4.39	3.38	0.59
	(1.33)	(1.16)	(1.74)	(0.95)	(1.15)	
Absence Rate at Endline (%)	9.09	9.72	10.50	9.99	8.10	0.72
	(2.09)	(2.04)	(2.18)	(1.63)	(2.00)	
Panel D: School Variables (N=49)				Program Schools	Pure Control Schools	p-value
# Classes Grade 3–6				5.48	6.25	0.32
				(0.43)	(0.76)	
Computer Lab				0.79	0.75	0.73
				(0.08)	(0.13)	
Local Population Density				0.18	0.19	0.63
				(0.01)	(0.02)	

Notes: This table presents the mean and standard error of the mean (in parenthesis) for several characteristics of students (Panels A & B), class rooms (Panel C), and schools (Panel D), across treatment groups. The student sample consists of all students tested by the research team during the baseline survey in February 2018. Column 6 shows the p-value from testing whether the mean is equal across all treatment groups. IRT-scores are standardized such that $\mu = 0$ and $\sigma = 1$ for the pure control group. The household asset index measures what share of the following assets a household owns: Books, electricity, television, washmachine, computer, internet and car. Local population density is the municipality's population density measured in 1000 inhabitans per km^2. Standard errors are clustered at the class level in Panels A & B, and at the school level in Panel C. * p<0.10, ** p<0.05, *** p<0.01.

through a careful mapping to the national curriculum and a feedback loop involving the regional Ministry of Education and local education experts. The math problems presented to the children mostly required a written answer (as opposed to a multiple choice format) and were inspired by El Salvador's official textbooks as well as various international sources of student assessments. Section B in the appendix explains the design of our assessments step by step.

In the appendix, we further present detailed statistics on the distribution of student test scores and the difficulty of the items. Top or bottom coding is neither an issue with respect to students nor the selected items: Table B.1 shows that virtually all items (except one for fifth graders in the endline assessment) were at least once answered correctly or incorrectly. Likewise, Table B.2 documents that only about 0.5% of test-takers scored zero points, while nobody achieved the maximum score. In general, the assessments seem to nicely capture the different performance levels, with the scores being roughly normally distributed around a median of 30% (3rd graders) to 40% (6th graders) correct answers (see Figure B.2).

A particularly nice feature of our math assessments is that they allow us to project all outcomes on a common ability scale by drawing on *Item Response Theory (IRT)*(e.g. de Ayala, 2009). This implies that we can directly compare children across grades and express their learning gains between the baseline and the endline assessment in terms of how many additional school years would be required to reproduce the same effect. The conversion of our estimates into program effects measured in terms of additional school years is explained in the appendix B.

3.2.2 Socio-Demographic Survey

The socio-demographic survey was distributed about 15 minutes before the baseline math assessment began. It asked students about their age, gender, household composition, household assets and parental education. Since literacy can be an issue, questions were illustrated with pictures and the enumerators helped children to understand and answer them correctly.

3.2.3 Administrative Data on Schools

In the run-up to the study, we collected various administrative data on Morazán's schools. While the government gathers vast information on the school environment through a paper-and-pencil survey administered to school principals, the data turned out to be of rather poor quality. To obtain usable information on the class structure, enumerators had to call each school at the beginning of the school year because the planning data from official sources was too unreliable. Moreover, the paper-and-pencil surveys left many missing values, so that we had to discard most items due to an insufficient coverage. We therefore decided to use a minimal set of school level variables, which were either comprehensively available, relatively cheap to supplement, or essential for the study. These include the number of grade three to grade six classes (i.e. school size), information on the presence of gangs (i.e. security at school), accessibility measures based on Google-Maps estimates and validated by local staff, power supply according to the administrative survey and validated via phone calls, student access to computer labs according to the administrative survey and validated via phone calls, and local population density from the National Bureau of Statistics.

3.2.4 Monitoring Data

From May to September 2018, NGO staff made on average five unannounced school visits (about 1000 visits in total) to collect monitoring data. These visits covered regular lessons in *pure control schools* as well as both regular and treatment lessons in *program schools*. The enumerators collected data on teacher attendance, student attendance, computer usage, and the implementation of the additional math lessons in the afternoon.

4 Results

4.1 The Overall Program Effects

We begin by estimating *intent to treat* (ITT) effects of being assigned to one of the three programs (i.e. $\beta_{T1}, \beta_{T2}, \beta_{T3}$) or the program school control classes (i.e. β_{CX}) using

$$Y_{icsk}^{EL} = \alpha + \beta_{T1} T1_{csk} + \beta_{T2} T2_{csk} + \beta_{T3} T3_{csk} + \beta_{CX} CX_{csk} + \delta Y_{icsk}^{BL} + X_{icsk}^{'} \gamma + V_{csk}^{'} \lambda + \phi_k + \epsilon_{1icsk}, \quad (1)$$

where Y_{icsk}^{EL} is the endline math score of student i in class c, school s and stratum k; math scores are either measured as percentage of correct answers or as the IRT-score normalized to μ=0 and σ=1 based on the baseline score of the pure control group. The binary treatment indicators are defined as follows: $T1$ equals one for those assigned to extra math lessons conducted by a teacher, $T2$ equals one for those assigned to extra CAL lessons overseen by a supervisor, $T3$ equals one for those assigned to extra CAL lessons instructed by a teacher, and CX equals one for those assigned to program school control classes that are potentially subject to externalities. Our control variables include Y_{icsk}^{BL} which stands for the baseline math score, X_{icsk} representing a set of student-level control variables (i.e. age normalized by the average age at the same grade level, gender, household size and household assets), and V_{csk} comprising a set of classroom-level variables (i.e. indicator for grade level, class size and teacher gender). Finally, ϕ_k stands for k strata fixed effects and ϵ_{1icsk} represents the error term.

The top panel of Table 2 displays the program effect relative to pure control classes (i.e. $\hat{\beta}_{T1}$, $\hat{\beta}_{T2}$, $\hat{\beta}_{T3}$ and $\hat{\beta}_{CX}$) and the grey-shaded rows of Table 2 present estimates for the pairwise differences between the three treatment groups within program schools. The within program school comparisons show p-values obtained from a randomization inference test statistic based on 2,000 random draws subject to the identical cut-off criterion as used in our re-randomization scheme (see section 3). The p-values for comparisons with pure control classes are based on school-level clustered standard errors, since the assignment to program schools and pure control schools did not involve re-randomization.[10] Density plots that provide a graphical representation of learning gains across different experimental conditions are shown in Figure A.1 in the appendix.

Students who were assigned to one of the treatments perform significantly better in the endline assessment than students assigned to the pure control classes. Compared to the pure control students, participants assigned to extra classes with math teachers (i.e. $T1$) perform 2.6 percentage points or 0.15σ better, students assigned to CAL classes with supervisors (i.e. $T2$) perform about 3.9 percentage points or 0.22σ better, and students assigned to CAL classes with a teacher (i.e. $T3$) perform 4.3 percentage points or 0.24σ better. Remarkably, students in control classes within program schools (i.e. CX) also score 2.4 percentage points or 0.14σ higher than students in pure control classes. As we discuss in section 5.1, our analysis points towards spillovers from CAL-lessons to program school control classes, while we find no evidence for direct exposure of control units or behavioral changes at the level of students, regular teachers or school administrators.

Finally, we test whether the observed gaps in the endline performance of students assigned to one of the three treatments (defined as β_4, β_5, and β_6) are statistically different from zero. While we find that the two CAL treatments outperform additional math classes, only the difference

[10]Moreover, we cannot properly apply randomization inference for comparisons with pure control classes due to missing information on ability levels of non-selected classes from pure control schools. As we show in appendix A.4, randomization inference for comparisons with pure control classes is based on draws that include on average 37% missing data points. Consequently, p-values obtained from these randomization tests increase by a factor of about 5 to 10 compared to p-values from traditional inference with school-level clustered standard errors. While this is clearly too conservative, our main conclusion are not altered when we apply randomization inference or comparisons with pure control classes (see Table A.3). The only notable difference is that program externalities, captured by β_{CX}, turn insignificant with p-values around 0.13. When we apply traditional inference to the within program school comparisons, as shown in Table A.2, changes in p-values are very small and do not show a clear pattern.

Table 2: ITT-Estimates on the effects of the different interventions on children's math scores

	Percent Correct		IRT-Scores	
	(1)	(2)	(3)	(4)
Comparisons with Pure Control Classes				
β_{T1}: Lessons with Teachers	2.904***	2.643**	0.165***	0.152**
	(0.005)	(0.016)	(0.006)	(0.016)
β_{T2}: CAL-Lessons with Supervisor	4.095***	3.869***	0.226***	0.214***
	(0.000)	(0.000)	(0.000)	(0.000)
β_{T3}: CAL-Lessons with Teacher	4.554***	4.328***	0.250***	0.238***
	(0.000)	(0.000)	(0.000)	(0.000)
β_{CX}: Control Classes for Externalities	2.595***	2.407**	0.147**	0.137**
	(0.009)	(0.020)	(0.011)	(0.021)
Comparisons within Program Schools				
$\beta_4 := \beta_{T2} - \beta_{T1}$	1.191	1.226	0.061	0.063
p-value (β_4=0)	(0.214)	(0.194)	(0.268)	(0.244)
$\beta_5 := \beta_{T3} - \beta_{T1}$	1.650*	1.686*	0.084	0.086
p-value (β_5=0)	(0.069)	(0.059)	(0.117)	(0.102)
$\beta_6 := \beta_{T3} - \beta_{T2}$	0.459	0.460	0.024	0.023
p-value (β_6=0)	(0.618)	(0.615)	(0.650)	(0.653)
Adjusted R^2	0.66	0.67	0.69	0.70
Observations	3197	3197	3197	3197
Individual & Classroom Controls	No	Yes	No	Yes
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

Notes: The mean (standard deviation) of the dependent variable for pure control units at *baseline*: Columns 1 & 2=31.14 (16.31); Columns 3 & 4=0.00 (1.00). At *endline*: Columns 1 & 2=34.46 (14.98); Columns 3 & 4=0.07 (0.98). The p-values for comparisons with pure control classes (coef. $\beta_{T1} - \beta_{CX}$) are based on school-level clustered standard errors. The p-values for comparisons within program schools (coef. $\beta_4 - \beta_6$) are based on a two-sided randomization inference test statistic that the placebo coefficients are larger than the actual; randomization inference is based on 2000 random draws. p-values are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

between additional math classes and CAL classes conducted by a teacher is statistically significant at the 10%-level: students assigned to CAL+TEACHER outperform students assigned to TEACHER by 1.7 percentage points or 0.085σ with p-values ranging from 0.059 to 0.117.

On the one hand, this is novel evidence that CAL delivers sizable learning gains in a Latin American context using off-the-shelf learning software: Expressing the estimates in terms of school years suggests that the effect of the CAL interventions is equivalent to the average student's progress in 0.6 to 0.7 school years (see appendix B for details on this conversion). On the other hand, the performance difference between CAL classes taught by teachers and additional teacher-centered math classes is statistically (marginally) significant. We interpret this as suggestive evidence that the learning gains reported in a series of CAL-evaluations can – at least partially – be attributed to the learning software and not necessarily to the increase in the number of math lessons.

4.2 Heterogeneity Analysis

We now examine effect heterogeneity along several dimensions. We first decompose program effects by subtopics, before we explore effect heterogeneity along baseline ability, grade level and class size.

4.2.1 Program Effects by Subtopic

In this subsection, we explore the impact of the three interventions on learning outcomes by topics. In accordance with the official curriculum, 65% of the items cover number sense and arithmetic (NSEA), 30% of the items cover geometry and measurement (GEOM), and 5% of the items cover data, probability and statistics (DSP). In particular, we re-estimate equation (1) but calculate separate math scores based on *(i)* NSEA-questions and *(ii)* GEOM- as well as DSP-questions.

The ITT-effects on students' NSEA skills are shown in columns (1) and (3) of Table 3. We find that both CAL treatments had a more pronounced effect on the NSEA score than on the overall math ability. Students who were assigned to CAL classes with supervisors score 4.6 percentage points or 0.24σ higher in NSEA questions than students assigned to pure control classes; this is an increase of 10% to 20% compared to the overall impact reported in Table 2. The NSEA score of students assigned to CAL classes with teachers is 4.9 percentage points or 0.26σhigher than the score of students in pure control classes; again this effect is 10% to 15% larger compared to estimates based on all questions. Since the impact on the NSEA score remains unchanged for students attending additional lessons instructed by teachers, the gap between CAL and conventional teaching widens.

Table 3: ITT-Estimates on the effects of the interventions on children's math score by subtopic

	Percent Correct		IRT-Scores	
	NSEA (1)	GEOM & DSP (2)	NSEA (3)	GEOM & DSP (4)
Comparisons with Pure Control Classes				
β_{T1}: Lessons with Teachers	2.849***	2.132*	0.146**	0.140**
	(0.010)	(0.067)	(0.017)	(0.040)
β_{T2}: CAL-Lessons with Supervisor	4.581***	3.014***	0.238***	0.187***
	(0.000)	(0.006)	(0.000)	(0.004)
β_{T3}: CAL-Lessons with Teacher	4.895***	3.472***	0.259***	0.193***
	(0.000)	(0.005)	(0.000)	(0.007)
β_{CX}: Control Classes for Externalities	2.463***	2.561**	0.130**	0.149*
	(0.009)	(0.048)	(0.015)	(0.054)
Comparisons within Program Schools				
$\beta_4 := \beta_{T2} - \beta_{T1}$	1.732*	0.882	0.091	0.047
p-value (β_4=0)	(0.093)	(0.432)	(0.115)	(0.464)
$\beta_5 := \beta_{T3} - \beta_{T1}$	2.047**	1.340	0.112*	0.053
p-value (β_5=0)	(0.047)	(0.221)	(0.055)	(0.412)
$\beta_6 := \beta_{T3} - \beta_{T2}$	0.315	0.458	0.021	0.006
p-value (β_6=0)	(0.752)	(0.669)	(0.714)	(0.926)
Adjusted R^2	0.63	0.47	0.65	0.50
Observations	3197	3197	3197	3197
Individual & Classroom Controls	Yes	Yes	Yes	Yes
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

Notes: *NSEA*=number sense & elementary arithmetic; *GEOM*=geometry & measurement; *DSP*=data, statistics & probability. The p-values for comparisons with pure control classes (coef. $\beta_{T1} - \beta_{CX}$) are based on school-level clustered standard errors. The p-values for comparisons within program schools (coef. $\beta_4 - \beta_6$) are based on a two-sided randomization inference test statistic that the placebo coefficients are larger than the actual; randomization inference is based on 2000 random draws. p-values are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

When we compare the learning gains attributed to CAL with the gains attributed to the additional math classes without software the differences range from 1.7 to 2.0 percentage points or from 0.091σ to 0.112σ. The corresponding p-values are 0.047 and 0.055 for the CAL classes with teachers and 0.093 and 0.115 for CAL classes with supervisors. Hence, when focusing on NSEA questions, the overall pattern remains qualitatively similar to the estimations including all subject domains, but the gap between the two CAL treatments and additional math classes in the traditional sense (i.e. without the use of software) becomes more pronounced.

Columns (2) and (4) of Table 3 show the results that are based on GEOM- and DSP-items. Focusing on these topics reduces the impact of both CAL treatments. The effects compared to pure control classes remain significant but they decrease considerably in magnitude. The results show, for instance, that additional CAL lessons conducted by a teacher increase the NSEA-score by about 5 percentage points, while the increase in the combined GEOM- and DSP-score is only 3.5 percentage points. Since this drop is less pronounced for those classes receiving additional math lessons instructed by a teacher, the within treatment school comparisons yield insignificant effects.

These results show that computer-assisted learning software can be a valuable substitute to traditional teaching, but its impact seems to be sensitive to the concepts that are taught. While we obtain a consistently positive value-added of CAL+TEACHER and CAL+SUPERVISOR relative to TEACHER, the measured differences seem to be primarily driven by the pronounced improvements in the domains of number sense and elementary arithmetic. The CAL interventions were less successful in shifting abilities to solve questions on geometry, measurement, data and statistics: The difference in point estimates decreases by about 20%, and the p-values clearly exceed the 0.1-threshold for statistical significance.

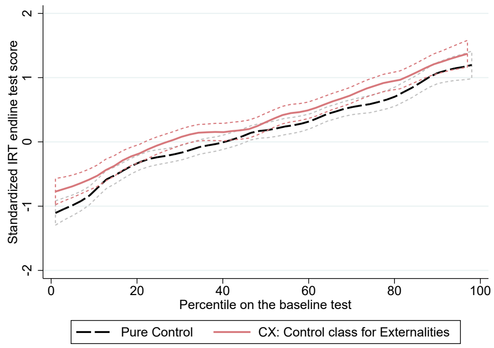
The noticeable differences across domains also suggest that CAL may not be well-equipped to substitute for all aspects of the complex task a teacher is expected to perform. While it may be relatively easy to automate the correction of errors in simple arithmetic exercises, evaluating students' progress and providing helpful feedback on tasks that require creativity or connected thinking may be much harder for a computer. Moreover, CAL may face a difficult job in connecting instructed concepts to real world experiences. While a teacher can, for example, distribute rulers to make students measure different objects in the classroom, pure CAL instruction is limited to what can be achieved with a two-dimensional screen. This suggests that a blended learning approach, where CAL is combined with active teachers who focus their engagement on tangible tasks may be a promising way to go.

In our setting, however, teachers did not compensate for potential weaknesses of the software in conveying concepts in geometry and statistics. The decline in point estimates between columns (3) and (4) of Table 3 for treatment CAL+TEACHER is about equivalent to the drop for treatment CAL+SUPERVISOR. While this pattern, similar as our other findings, yields little indication for a vital complementarity between CAL software and teachers, further evidence is required to generalize this conclusion to other contexts. The degree of complementarity likely depends on the concepts taught, the design of the software, and the ability as well as mandate of the instructor. Considering that the role of the instructor is a central and costly ingredient to the implementation of a CAL initiative, further research on the interplay between instructors and software may prove valuable.

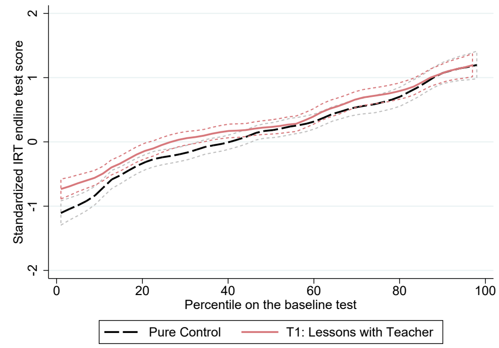
4.2.2 Effect Heterogeneity by Baseline Ability, Grade Level and Class Size

We continue the heterogeneity analysis by discussing Figure 3, which plots kernel-weighted locally-smoothed means of the endline test score at each percentile of the baseline test score by treatment status. Figure 3a shows that endline tests scores in the control group for spillovers are slightly higher than those in the pure control group at all percentiles of the baseline test score, but the 95% confidence bands mostly overlap. Comparing pure control classes to the TEACHER classes in Figure 3b shows that the latter outperform the former at low percentiles of the baseline score, while there is no difference at higher percentiles. Both CAL intervention groups, as illustrated in Figures 3c and 3d, achieve considerably higher endline scores than pure control classes across all percentiles in the baseline achievement, although the gap seems to narrow at higher percentiles in the CAL+TEACHER group.

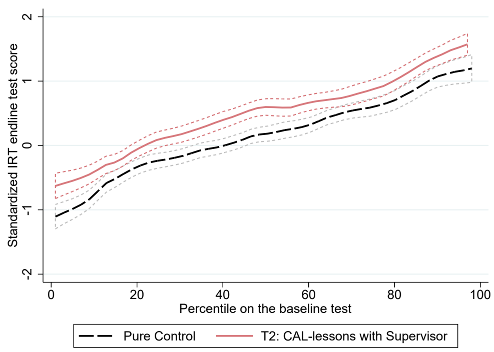
In a next step, we examine the functional relation between treatment effects and baseline achieve-



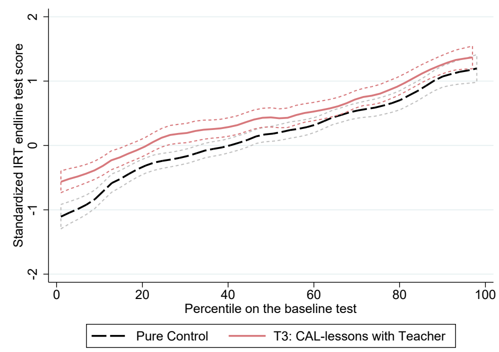
(a) Spillover vs. pure control classes



(b) TEACHER vs. pure control classes



(c) CAL+SUPERVISOR vs. pure control classes



(d) CAL+TEACHER vs. pure control classes

Figure 3: Endline test scores by treatment status and baseline percentiles.
Note: The figures present kernel-weighted local mean smoothed plots relating endline test scores to percentiles in the baseline achievement by treatment status alongside 95% confidence bands.

ment more thoroughly. We further investigate whether the reported effects vary by grade level or by class size. To do so, we estimate

$$\begin{aligned}
Y^{EL}_{icsk} = \alpha &+ \beta_{T1}T1_{csk} + \beta_{T2}T2_{csk} + \beta_{T3}T3_{csk} + \beta_{CX}CX_{csk} \\
&+\theta_{T1}(T1_{csk} \times Var_{icsk}) + \theta_{T2}(T2_{csk} \times Var_{icsk}) \\
&+\theta_{T3}(T3_{csk} \times Var_{icsk}) + \theta_{CX}(CX_{csk} \times Var_{icsk}) \\
&+\delta Y^{BL}_{icsk} + X^{'}_{icsk}\gamma + V^{'}_{csk}\lambda + \phi_k + \epsilon_{2icsk}
\end{aligned} \tag{2}$$

where $(T_{csk} \times Var_{icsk})$ is the interaction of the treatment dummy with the variable of interest (i.e. baseline math score, grade level and class size). Except for the four interaction terms, equation (2) is identical to our benchmark estimation equation, i.e. equation (1).

Table 4: Effect heterogeneity along baseline ability, grade level and class size

Interacted with:	Baseline Math Score	Grade Level	Class Size (log)
Dep. Var.: IRT Score	(1)	(2)	(3)
Main Effects			
β_{T1}: Lessons with Teachers	0.149**	0.151**	0.133**
	(0.015)	(0.014)	(0.025)
β_{T2}: CAL-Lessons with Supervisor	0.214***	0.216***	0.204***
	(0.000)	(0.000)	(0.000)
β_{T3}: CAL-Lessons with Teacher	0.240***	0.240***	0.227***
	(0.000)	(0.000)	(0.000)
β_{CX}: Control Classes for Externalities	0.139**	0.143**	0.127**
	(0.022)	(0.022)	(0.024)
Interaction Terms (Var.=Column Header)			
θ_{T1}: Lessons with Teachers × Var.	−0.105***	−0.140***	−0.436***
	(0.002)	(0.000)	(0.004)
θ_{T2}: CAL-Lessons with Supervisor × Var.	−0.014	−0.051	−0.110
	(0.750)	(0.263)	(0.400)
θ_{T3}: CAL-Lessons with Teacher × Var.	−0.038	−0.058	−0.270**
	(0.284)	(0.236)	(0.032)
θ_{CX}: Control Classes for Extern. × Var.	−0.005	−0.022	−0.118
	(0.912)	(0.711)	(0.442)
Adjusted R^2	0.70	0.70	0.70
Observations	3197	3197	3197
Individual & Classroom Controls	Yes	Yes	Yes
Baseline Score	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes

Notes: All interaction variables are demeaned. p-values are based on school-level clustered standard errors and are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

In terms of baseline math ability, the regression analysis confirms our visual analysis of Figure 3. Regarding the effect of additional math classes instructed by teachers, the effect size and baseline achievement are indeed negatively correlated (see column 1 in Table 4). This suggests that teachers were more effective in improving the performance of children with low math ability than those children who performed well in the baseline assessment. The regression also yields negative signs for the interaction between the baseline math score and T2 (i.e. CAL+SUPERVISOR) as well as T3 (i.e. CAL+TEACHER), but the p-values do not reach the 10%-threshold (T2: p-value=0.75; T3: p-value=0.28). If we evaluate the impact of each treatment for students one standard deviation above the mean baseline ability, we obtain effect sizes of 0.05 (p-value=0.50) for TEACHER, 0.20 (p-value<0.01) for CAL+SUPERVISOR, and 0.20 (p-value<0.01) for CAL+TEACHER. Hence, the benefit of attending CAL-based lessons was independent of initial ability levels, while the effectiveness of teachers without software was particularly low among well-performing students.

A similar pattern emerges when we study effect heterogeneity by grade level of the participating students (see column 2 in Table 4). The effects of the CAL treatments do not significantly vary along grade level of students, but we find that additional lessons taught by a teacher are less effective in higher grades. Estimating the models per grade yields noisy estimates as each treatment contains only between 9 to 11 classes per grade. The grade specific analysis emphasizes the pattern reported in column 2: The gap between TEACHER and CAL+TEACHER widens to 0.20σ (p-value=0.09) for grade five and 0.22σ (p-value=0.10) for grade six. Similarly, the difference between TEACHER and CAL+SUPERVISORS increases to 0.22σ (p-value=0.06) for grade five and 0.12σ (p-value=0.36) for grade six.[11] This corroborates the finding that without the help of CAL software, teachers in Morazán are least effective when explaining more complex concepts.

Finally, we find that large class sizes reduce the effectiveness of teachers (see column 3 in Table 4), no matter whether they use CAL software or not. This pattern does not emerge for CAL classes overseen by supervisors, which seems plausible since supervisors were directed to refrain from explaining math contents but solely provided technical assistance. Comparing the point estimates of the interaction terms of the two treatments conducted by teachers, we find that the effect of traditional classes ($\hat{\theta}_1$=0.44, p-value<0.01) is considerably more sensitive to class size than the effect of CAL-lessons instructed by teachers ($\hat{\theta}_3$=0.27, p-value=0.03). If we evaluate the impact of

[11] This also reflects in the impact estimates measured as school year equivalents: For the TEACHER treatment, the ITT-estimates correspond to about 1 school year for grade three and grade four, but essentially 0 for grade six and even negative for grade five. In case of both CAL treatments the ITT-estimates fluctuate around 1 school year, except for grade five where the estimates decrease to an equivalent of about 0.4 school years.

each treatment for students attending classes with a size of one standard deviation above the mean (i.e. 27 students), we obtain effect sizes of -0.01 (p-value=0.87) for TEACHER, 0.17 (p-value=0.01) for CAL+SUPERVISOR, and 0.14 (p-value=0.04) for CAL+TEACHER. Overall, this strongly confirms the notion that computer-based learning can mitigate the problems related to large classes (e.g. Banerjee and Duflo, 2011; Muralidharan et al., 2019).

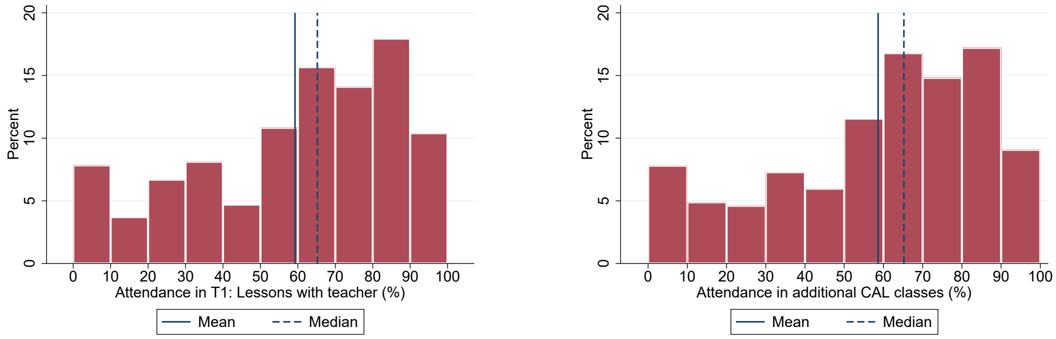
4.3 Program Attendance and IV-Estimates

Our benchmark analysis focuses on ITT-estimates that do not account for the actual attendance rate of students in the additional math lessons. In this section, we present data on the overall compliance, examine the correlation between individual attendance and endline scores, and finally discuss instrumental variable estimates for the impact of the three interventions assuming full attendance.

Figure 4 plots the distribution in attendance rates across all eligible students. With an average attendance rate of 59%, participation of students was a weak spot of the program. Attendance rates slightly varied across the three treatments, although the differences are statistically insignificant: Additional CAL classes instructed by teachers achieved the highest participation (60%), followed by additional classes instructed by teachers (59%) and CAL classes conducted by a supervisor (57%). Across all treatments, average participation started at 65% and then gradually declined to 55% over the course of the program. As Figure 4 shows, about 7% of the students did participate in less than five lessons, while the median student attended roughly two thirds of the classes.[12]

The individual attendance rate of students is strongly correlated with their performance in the endline math assessment, as one would expect considering that the programs successfully increased math learning outcomes.

[12] One key challenge to participation was related to safety concerns. As there are no afternoon buses to many remote communities, some parents were worried about students having to cross dangerous areas by foot. Moreover, some households rely on children's contributions to the family business or household in the afternoons.



(a) T1: Additional math lessons with teachers

(b) T2 & T3: Additional CAL-lessons

Figure 4: Attendance of students in additional math lessons.

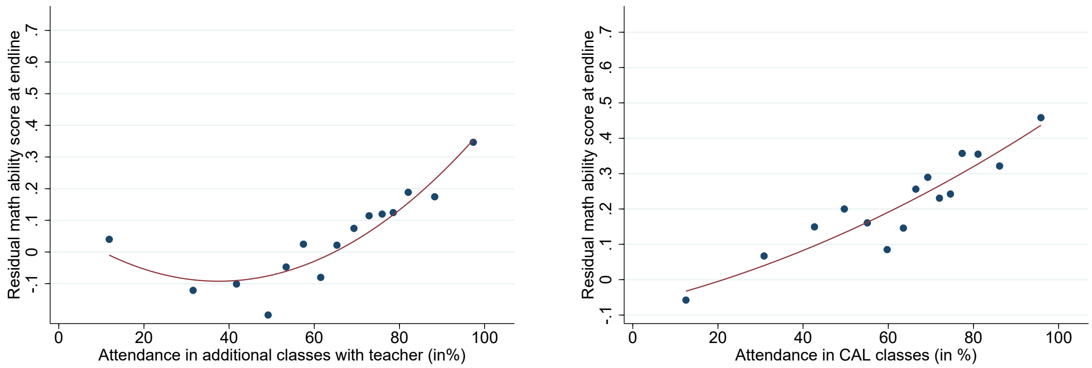
Figure 5 plots the residual endline IRT-score (net of all control variables including baseline scores) on the y-axis, and the attendance rates of the students on the x-axis. We aggregated the individual data points into 15 bins in order to improve readability, and plot the correlation by treatment type. Figure 5a covers the students that were assigned to additional math classes taught by teachers, while Figure 5b illustrates the correlation between attendance and residual endline scores for the two CAL interventions.[13]

We next appraise the question, how much children would have learned had they fully participated in the additional math lessons they were offered. To do so, we estimate an IV-model, with the first-stage estimation being specified as

$$Att_{icsk}^{T=t} = \alpha + \pi_1 T1_{csk} + \pi_2 T2_{csk} + \pi_3 T3_{csk} + \delta Y_{icsk}^{BL} + X_{icsk}^{'}\gamma + V_{csk}^{'}\lambda + \phi_k + \epsilon_{3icsk} \quad for \quad t \in [1,2,3] \tag{3}$$

where $Att_{icsk}^{T=t}$ is student's i attendance rate in treatment t and takes values between 0 and 1.

[13]Regressing endline IRT scores on attendance rates (continuous between 0 and 1), baseline scores, individual and classroom controls yields the following correlations between attendance and performance: $\hat{\gamma}_{T1}$=0.40 (t-value=5.0); $\hat{\gamma}_{T2}$=0.56 (t-value=4.2); $\hat{\gamma}_{T3}$=0.55 (t-value=3.6). Including a quadratic term we get: $\hat{\gamma}^1_{T1}$=−0.53 (t-value=−1.9), $\hat{\gamma}^2_{T1}$=0.89 (t-value=3.1); $\hat{\gamma}^1_{T2}$=0.56 (t-value=1.1), $\hat{\gamma}^2_{T2}$=0.01 (t-value=0.0); $\hat{\gamma}^1_{T3}$=−0.41 (t-value=−1.0), $\hat{\gamma}^2_{T3}$=0.94 (t-value=2.1).



(a) T1: Additional math lessons with teachers (b) T2 & T3: Additional CAL-lessons

Figure 5: Residual endline test scores and attendance in additional math lessons.
Note: The figures present the partial correlation between individual attendance rates and residual endline test scores after controlling for baseline scores, individual and classroom characteristics. To ease readability, we aggregated individual data points into 15 bins.

All other variables are defined as in the benchmark estimation equation, i.e. equation (1). In the second stage, we replace the binary treatment indicators with the predicted attendance rates from stage 1, i.e. $\widehat{Att}_{icsk}^{T=t}$, and estimate

$$Y_{icsk}^{EL} = \alpha + \beta_{T1}^{IV}\widehat{Att}_{icsk}^{T=1} + \beta_{T2}^{IV}\widehat{Att}_{icsk}^{T=2} + \beta_{T3}^{IV}\widehat{Att}_{icsk}^{T=3} + \delta Y_{icsk}^{BL} + X_{icsk}^{'}\gamma + V_{csk}^{'}\lambda + \phi_k + \epsilon_{4icsk}. \quad (4)$$

In order to interpret $\hat{\beta}_{T1}^{IV}$, $\hat{\beta}_{T2}^{IV}$, and $\hat{\beta}_{T3}^{IV}$ as the treatment effects of attending all 46 additional math lessons, we have to impose two (restrictive) properties that go beyond the standard monotonicity and independence assumptions (see Angrist and Pischke, 2008; Muralidharan et al., 2019). *First*, the treatment effect needs to be homogenous across students. *Second*, the functional form between attendance and math score gains has to be linear.

Our data suggest that these two additional assumptions may be violated and that the IV-estimates are potentially *downward* biased. *Effect homogeneity* seems questionable, since the impacts of the interventions are homogenous (in case of both CAL treatments) or decreasing (in case of the TEACHER treatment) in initial ability, even though attendance rates are positively correlated with baseline scores. Attending an additional math lesson thus had a stronger effect on low

Table 5: IV-Estimates: Program effects with full participation

	Percent Correct		IRT-Scores	
	(1)	(2)	(3)	(4)
β^{IV}_{T1}: Lessons with Teachers	5.066***	4.739***	0.286***	0.269***
	(0.001)	(0.003)	(0.001)	(0.003)
β^{IV}_{T2}: CAL-Lessons with Supervisor	7.104***	6.859***	0.390***	0.378***
	(0.000)	(0.000)	(0.000)	(0.000)
β^{IV}_{T3}: CAL-Lessons with Teacher	7.517***	7.236***	0.411***	0.396***
	(0.000)	(0.000)	(0.000)	(0.000)
Kleibergen-Paap F-statistic	99.78	92.34	99.93	92.44
Adjusted R^2	0.65	0.66	0.69	0.69
Observations	2570	2570	2570	2570
Individual & Classroom Controls	No	Yes	No	Yes
Baseline Score	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes

Notes: p-values based on school-level clustered standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

ability than high ability students. Hence, the IV-estimates might understate the true effect under full participation. Moreover, the functional form between attendance and ability gains appears to be (slightly) convex rather than linear, suggesting that children experienced increasing returns to attending the additional lessons. Again this would lead to a downward bias in the IV-estimates.

Table 5 presents the IV-estimates, which can be interpreted as the (potentially downward biased) treatment effects of attending all 46 additional math lessons. Attending the full CAL program during the intervention period leads to an increase in the endline score of about 7 percentage points or 0.38σ to 0.41σ, which is comparable in magnitude to effects of technology-aided instruction found in India, where Muralidharan et al. (2019) report average learning gains in math of 0.6 standard deviations for 90 days attendance at CAL learning centers.

An increase in math ability of 0.4σ is about equivalent to the average student's progress in 1.1 school years. This translation of the average treatment effects under full compliance into school-year equivalents should be read with caution, however: On the one hand, school year equivalents do not only represent what children learn in their regular math classes at school but also reflect age-based cognitive development, learning at home or spillovers from other subjects. On the other hand, our monitoring data suggest that about 25% of regular lessons are canceled due to teacher absenteeism and that children miss another 10%. Hence, compliance in regular classes is far from perfect, which complicates statements concerning the relative effectiveness of the additional CAL-based lessons compared to regular math lessons based on these estimates.

5 Discussion

5.1 Treatment Externalities

Our research design allows us to quantify spillovers on non-treated classes in program schools. As discussed in section 4.1, we find positive and significant externalities: Students assigned to control classes in program schools scored about 0.14σ higher in the endline assessment than students assigned to pure control classes. This effect is comparable in magnitude to the treatment effect for additional math lessons instructed by teachers. While we do not have rigorous experimental evidence to pin down the mechanisms with certainty, the data we collected from various sources allows for an assessment of the plausibility of different channels. In the following we distinguish between three broad explanations: *(i)* direct exposure of students in control classes, *(ii)* behavioral adjustments to the experimental design, and *(iii)* social learning among peers.

Direct Exposure. We begin by examining the hypothesis that control students in program schools may have been directly exposed to one of the treatments, either by (illicitly) participating in the additional math lessons, by targeted migration and class changes, or by using CAL-software in regular lessons or at home.

To prevent direct exposure of control students to the treatments, the implementing NGO instructed contract teachers and supervisors to confine access to children that were registered as participants. Our monitoring data shows high compliance with this directive, as unauthorized participation was only recorded during 6 out of about 750 unannounced visits in treatment classes.

Likewise, we aimed to eliminate any incentives to change classes or schools and therefore barred students that changed into treatment classes during the school year from attending the additional math lessons. Only 38 students in our estimation sample changed classes or schools during the program and excluding these students from the estimation models leaves the results unchanged.

Control students in program schools may also have been exposed to the learning software in regular classes or at home. To prevent this, the implementing NGO stored the supplied hardware in lockable cabinets and used password-protected logins. While regular teachers may have been inspired to set up their own CAL facility or control students may have used learning software at home, our data suggests that this is unlikely. Enumerators recorded computer usage in only 5 out of about 1,000 regular class visits, and less than 20% of children live in a household owning a computer with internet access. Moreover, neither the students' ownership of a computer with internet access at home nor internet access at the school level are significantly correlated with learning outcomes in the endline assessment.[14]

In line with this reasoning, the domain-specific treatment externalities presented in Table 3 also contradict the hypothesis that control units in program schools were directly exposed to the learning software. While the CAL treatment effects are strongest in the domain for number sense and elementary arithmetic, the treatment externalities do not match this pattern but accrue equally across domains.

[14] To further assess whether the observed externality may operate via online usage of learning software, we interacted the binary indicator for control classes subject to externalities with a binary indicator for internet access; neither students' internet access at home ($\hat{\beta}$=0.01, pvalue=0.82) nor internet access at the school-level ($\hat{\beta}$=0.09, pvalue=0.55) yield a significant estimate for this interaction term.

Behavioral Adjustments to the Experimental Design. There are several ways in which students or school staff might have inadvertently adjusted their behavior to the experiment (see appendix A.5 for a more detailed discussion). *First*, the presence of the NGO could have *incentivized* schools to make a good impression, for instance to be allowed to keep the IT equipment after the intervention or to be considered for future collaborations. Our data does not support this claim, however, as neither teacher nor student attendance is higher in program schools (see Table A.4) and the number of installed computers is uncorrelated with student performance (see Table 6).

Second, the difference between control classes within and outside program schools may be driven by a *John Henry effect*, a bias induced from reactive behavior of the control group to overcome the *disadvantage* of not being treated (e.g. Glennerster and Takavarasha, 2013). If such behavior arises within program schools, but not in spatially separated pure control schools, it could account for the observed externalities. This mechanism has similar implications, but is distinguishable from those discussed in the previous paragraph. While the previous paragraph explores the possibility of a general boost in student or teacher motivation across all groups in treatment schools, the John Henry effect would only operate for the control group. As shown in columns (3) and (6) of Table A.4, the data does not support this hypothesis.

Third, the experiment might have induced behavioral changes in response to being *observed*, so called *Hawthorne effects* (e.g. Levitt and List, 2011). If being part of an experiment was more salient to subjects in program schools, they might have worked harder as a response to being observed, producing the pattern we find in our data. This seems unlikely for several reasons. Most importantly, the monitoring process was structured homogeneously, meaning that enumerators visited program and control schools with the same frequency and followed the same observational procedure. Moreover, only few studies provide evidence for the presence of Hawthorne effects in the context of educational interventions, even though the topic received considerable attention (e.g.

Table 6: Externality channel: Proxies for social learning and in-kind incentives

Dep. Var.: IRT Score	Treatment Intensity		Installed NGO Computers	
CX-Indicator Interacted with:	All Treatments	CAL	Per Student	Total
	(1)	(2)	(3)	(4)
Control Classes for Externalities	0.146**	0.135**	0.142**	0.146**
	(0.019)	(0.023)	(0.020)	(0.037)
Control Classes for Externalities × Var.	0.010	0.015***	0.031	0.001
	(0.290)	(0.001)	(0.950)	(0.865)
Adjusted R^2	0.73	0.74	0.73	0.73
Observations	1279	1279	1279	1279
Individual & Classroom Controls	Yes	Yes	Yes	Yes
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

Notes: Treatment intensity is defined as the percentage share of treated students in a school. All interaction variables were demeaned. p-values are based on school-level clustered standard errors and are shown in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Adair, 1984; Adair et al., 1989; Krueger, 1999).

Finally, divergent behavioral responses by treatment status might only have occurred during the math assessment. A large body of literature shows that *test-taking motivation* can have profound effects on low-stakes test results (e.g. Silm et al., 2020). Since the implementation of our experiment did not hint at any personal or institutional rewards for participants, it seems unlikely that the treatment status systematically influenced students' test-taking effort. One may further hypothesize that students in program schools put more effort into the tests because they perceived it as more "purposeful" when other classes of the same grade also participated in the examinations. We test this claim by interacting the control classes with a binary indicator equaling one for classes in schools that have other classes of the same grade level that took the test (almost exclusively satisfied in program schools), but do not find a significant correlation. Finally, even if motivational effects were present, one would expect them to have influenced performance during the baseline assessment as well, which would cancel out any potential bias operating via this channel.

Peer effects. The treatment externalities may also stem from *peer effects*, as participants could have shared their knowledge and/or elevated enthusiasm for math with schoolmates from other classes. Results in columns (1) and (2) of Table 6 suggest that this may have been the case: What explains part of the performance differential between program school control classes and pure control classes is the share of children that participated in the CAL treatments. One explanation is that the learning gains produced by CAL were passed on by the participants to their peers from non-treated classes. Another explanation for this pattern would be that hosting many CAL classes went along with a more generous furnishing of computer-labs by the NGO, which might have incentivized school staff to make a good impression with the NGO so that they could keep the equipment even after the NGO-run program expired. As discussed above, columns (3) and (4) in Table 6 show no relevant correlation between the number of NGO computers installed in a school and the endline performance of students. Hence, the interpretation that CAL beneficiaries passed on their learning gains and/or elevated enthusiasm for math to their peers seems more plausible than behavioral adjustments in prospect of being donated equipment. This finding is consistent with a literature of peer-effects that documents how the performance of each student affects achievements of her classmates (see Sacerdote, 2011).

Peer effects may also operate between regular teachers and contract teachers, for instance, if regular teachers imitate the techniques and materials used by contract teachers in the treatment lessons. Unfortunately, we did not monitor teaching methods in regular classes (beyond computer usage), so that we cannot examine this specific mechanism. From a program effectiveness point view, peer effects between teachers would hold very similar implications to peer effects among students.

Summarizing Remarks. Although we cannot comprehensively pin down the channels through which the observed externalities operate, *social learning among peers* is the mechanism that can be reconciled best with the data at hand. In contrast, we are confident to rule out *direct exposure* of control units to the evaluated treatments, as our data documents excellent compliance with the experimental protocol. *Behavioral adjustments to the experimental design* may unfold in many ways, which makes it difficult to track them exhaustively. We tested several potential channels of this category, but the data consistently rejects the underlying hypotheses.[15]

5.2 Cost-Effectiveness

Since all three interventions were assessed within the same context and framework, we can directly compare their cost-effectiveness. The bulk of expenditures comes from salaries to teachers and supervisors (65% for TEACHER, 41% for CAL+SUPERVISOR, and 51% for CAL+TEACHER). The two computer treatments additionally entail costs for acquiring the IT equipment, shipping it to El Salvador and maintaining it. Since our partner NGO acquired most computers as in-kind donations, the factual IT-related costs incurred by the NGO (about 18 USD per computer) provide a poor guidance for educational policy-makers aiming to implement CAL interventions at scale. To make

[15]The discussed channels imply three competing interpretations of our results: *First*, the program may have unfolded genuine externalities due to social learning among peers. This is supported by our data as we indeed find a correlation between the number of students attending CAL lessons and the performance of their peers in control classes. *Second*, the observed pattern may not result from actual externalities, but from a biased estimate for the program school control classes caused by John Henry effects. John Henry effects may operate via test effort on the assessment day (not testable with our data) or throughout the implementation phase (no supporting evidence in our data), but – in either case – they would *not* affect the interpretation of the three treatment estimates. A *third* possibility is that the pure control classes do not constitute a sound counterfactual because students from program and control schools differed systematically in their test taking effort. In this the case, the performance differential between the two control groups would have to be subtracted from the impact estimates for each treatment (see appendix A.6), roughly halving the impact the two CAL treatments ($T2$: p-values=0.10-0.13; $T3$: p-values=0.04-0.06) and virtually eliminating the impact of the teacher treatment ($T1$: p-values=0.72-0.78). While we do not find any supporting evidence for this claim, we do not possess the data to rule it out with certainty. Since such behavior might occur in a large number of educational field experiments, incorporating measures of test taking effort in future experiments may yield important insights about this potential methodological artifact.

the cost-effectiveness calculations more insightful for a generic setting, we assume costs of 200 USD per work station and an average of five years of usage time.

Based on these assumptions for the costs of the computer hardware, the cost accounting of our partner NGO, and the guidelines developed by Dhaliwal et al. (2014), we estimate the cost per child to be 44 USD for TEACHER, 43 USD for CAL+SUPERVISOR, and 56 USD for CAL+TEACHER. Assuming a linear dose-response-relationship, TEACHER can thus be expected to yield a 0.35σ increase in test scores per 100 USD, while investing the same amount of money in CAL lessons would produce 0.49σ and 0.43σ, respectively. This implies that even when the computers have to be acquired at a considerable price, the two CAL interventions outperform additional teacher-led classes in terms of cost-effectiveness. Moreover, hiring lower-paid supervisors rather than officially certified teachers to conduct the CAL classes might be slightly more cost-effective, as supervisor were paid only about 60% of a teacher's wage. These conclusions should be interpreted with caution: Not only is precision impaired by the statistical uncertainty of our estimates, but the relative cost-effectiveness also depends on contextual factors such as local wages.

5.3 The Role of Teacher Ability

Multifaceted evidence derived in our analysis indicates issues concerning the preparedness of teachers. *First*, the heterogeneity analysis shows that the productivity of teachers declines as the complexity of concepts increases: The impact of the additional math lessons instructed by a teacher is decreasing in both the grade level as well as the baseline achievement of their students. *Second*, both CAL interventions outperform the additional math lessons instructed by teachers: The point estimates of the CAL interventions are consistently larger, and they neither decrease significantly with student baseline performance nor grade level. Hence, it appears that in our setting, learning software is more productive in teaching basic math than officially certified teachers, especially as the complexity of the content increases. *Third*, teachers do not seem to add much to the effect of computer-assisted learning lessons: The estimated impact for CAL lessons instructed by teachers is only marginally and insignificantly higher than that of CAL lessons conducted by supervisors (p-values around 0.6).

To better understand this pattern, we asked the instructors hired by the NGO to participate in a 90 minutes math assessment covering the primary school curriculum of grades two to grade six. Moreover, we administered the same assessment to a representative sample of regular math teachers of grade three to grade six classes allowing us to learn how the contract teachers compare to the regular teaching staff. Both assessments took place in late 2018 and thus after the end of the field experiment. Figure 6 illustrates the main insights from this assessment: Primary school math teachers in the department of Morazán do not master large parts of the content they are supposed to teach. The contract teachers hired by the NGO answered on average only 75% of the second and third grade questions correctly and this share declines to 54% for the sixth grade questions. Hence, even for the simplest questions, the average contract teacher does not meet the minimum proficiency of 80% correct answers as advocated by the World Bank (see Bold et al., 2017a; World Bank, 2018).

These insights raise the question whether the teachers hired for the intervention have a particularly low proficiency in math – which could explain why they are not part of the publicly employed teaching staff. Figure 6 suggests otherwise: Regular teachers performed considerably worse than the contract teachers, as they achieved on average only 56% correct answers on second and third grade questions and 30% on items pertaining to the sixth grade curriculum; the differences in means between contract and regular teachers are also statistically significant at the 1% level.[16]

[16] Note that the implementing NGO administered a very short math assessment in the hiring process in order to eliminate the least qualified candidates. Moreover, the hired teachers participated in several workshop to prepare them for the teaching assignment. Since the assessment reported in Figure 6 was conducted *after* the intervention

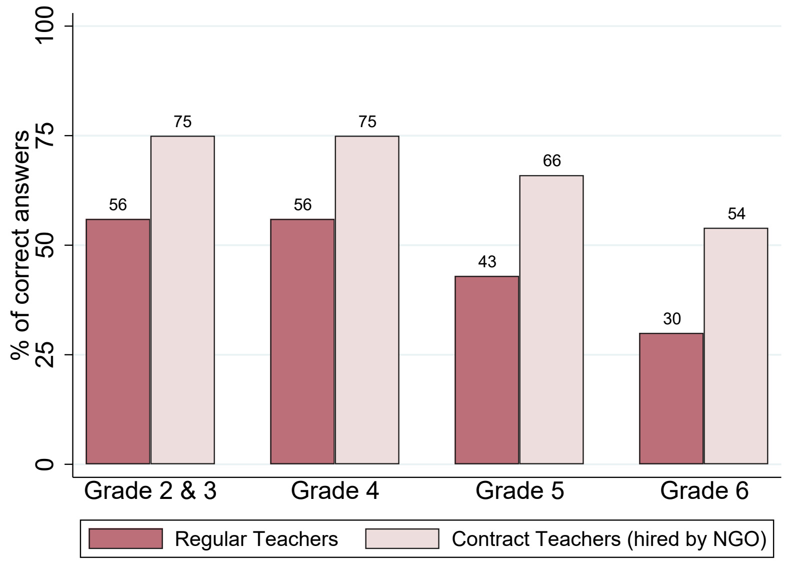


Figure 6: Math proficiency among regular teachers and teachers hired for additional math lessons.
Note: The graph shows the share of correct answers on questions covering the official math curriculum of grades 2 & 3, grade 4, grade 5, and grade 6. This data was collected after the endline assessment for students in late 2018 and early 2019. The sample includes all program teachers as well as a representative sample of regular primary school teachers teaching math in grades 3 to 6 in the department of Morazán. Contract teachers significantly outperformed regular teachers across items of all grades (p-values$<$0.01); the t-values for the plotted differences in means are 4.3 for items of grades 2/3, 3.3 for items of grade 4, 3.9 for items of grade 5, and 4.4 for items of grade 6.
Source: Brunetti et al. (2020).

Inadequate content knowledge of teachers likely puts a binding constraint on their productivity. To test this claim we conduct two additional analyses using the data on the math ability of contracted instructors as well as regular math teachers:[17] *First*, we re-estimate equation (2) and interact the three treatment dummies with the instructors' standardized math ability score. The math content knowledge of teachers is correlated with student learning gains in both traditional ($\hat{\theta}_1$=0.08, p-value=0.28) and CAL-based math lessons ($\hat{\theta}_3$=0.09, p-value=0.14), whereas the math

finished, it is likely that the NGO's selection process and the additional training for the contract teachers partly explains the pronounced differences in content knowledge between the regular teachers and the contract teachers.

[17]This additional analysis comes with some caveats: The experimental protocol did not take teacher ability into consideration, which is why teachers were assessed in the aftermath of the field experiment. Moreover, the number of different instructors was not optimized with respect to statistical power. The implementing NGO hired 23 teachers and 15 instructors to conduct the additional math lessons; all contract teachers instructed both traditional lessons (T1) and CAL-based courses (T3) so that an average contract teacher was responsible for two class per treatment arm.

score of supervisors is virtually orthogonal to learning gains ($\hat{\theta}_2$ <0.01, p-value=0.94). Since supervisors did not provide math related explanations, it makes sense that their math ability does not moderate the impact of CAL-based lessons. *Second*, we correlate the standardized math ability of regular teachers with the learning gains of their students between the baseline and the endline assessment. Depending on the model specification, the point estimates vary between 0.09 and 0.12 and are significant at the 0.05 level or higher (see Table 4 in Brunetti et al., 2020). Although these are purely observational estimates, they are not only very similar to the point estimates obtained for the impact of teacher ability in the program classes reported above, but also to quasi-experimental evidence established in several studies: The benchmark estimates for the annual impact of one standard deviation in additional teacher content knowledge on standardized learning outcomes of children are 0.09 for math in Peruvian primary schools (Metzler and Woessmann, 2012), 0.09 for math in Pakistani primary schools (Bau and Das, 2020), and 0.07 for math and language in Eastern African primary schools (Bold et al., 2019). Overall, these consistently positive point estimates for teacher content knowledge corroborate the hypothesis that the contract teachers' poor subject mastery impaired the impact of the evaluated teacher-centered math lessons. Hence, inadequate content knowledge is a plausible factor that helps to explain the comparatively low productivity of teachers reported in this study.

In view of drawing general conclusion for the effectiveness of additional math lessons instructed by regular teachers, the results reported in Figure 6 are particularly grim. The relatively low impacts found for the additional math lessons instructed by contract teachers may be too optimistic when aiming for a scale-up with regular teachers, who have on average a lower math proficiency than the contract teachers hired by the implementing NGO. According to our most reliable estimates, it would take a teacher with 88% correct answers on the administered assessment to teach additional math classes so that the attending students improve on average as much as students attending

CAL lessons overseen by a supervisor. This score corresponds to the 75th math ability percentile among the hired contract teachers and the 91st math ability percentile among regular primary school teachers in Morazán.[18]

These results highlight how learning software can compensate for the poor content knowledge of teaching staff. Earlier contributions on the value of computer-assisted learning emphasized its advantages in terms of mitigating issues of large class sizes and the challenges of "teaching at the right level" (e.g. Banerjee and Duflo, 2011; Muralidharan et al., 2019). While our heterogeneity analysis corroborates this line of reasoning, this section showed that CAL can help to remedy shortcomings related to low teacher ability.

6 Conclusion

Computer-assisted learning (CAL) is widely perceived as a promising approach to address the low quality of teaching in developing countries. While encouraging, previous research is inconclusive regarding the value of technology-based instruction relative to traditional teaching and and has little to say on the complementarities between teachers and learning software. The evidence presented in this paper suggests that CAL approaches can not only produce substantial learning gains, but may also outperform traditional instruction. At the same time, the results yield no indication for a substantive complementarity between CAL software and teachers.

The documented performance gap between CAL and teacher-centered lessons is likely driven by two aspects: *First,* the productivity of CAL-based instruction is less sensitive to class size, student ability, and grade level. *Second,* the field experiment is embedded in a schooling environment

[18] For this back-of-the-envelope calculation, we use 0.063σ for the impact difference between teacher-based additional lessons and CAL lessons monitored by supervisors (see Table 2), 0.09 for the conversion factor of standardized teacher content knowledge on standardized student ability, and data on content knowledge of teachers presented in Figure 6.

characterized by a mismatch between teacher content knowledge and the complexity of the concepts they have to teach. Under traditional teaching models, children are unlikely to learn what their teachers fail to understand, while CAL allows them to make progress beyond their teachers' content knowledge.

The observation that teachers struggle with many building blocks of the primary school curriculum may also explain our finding that teachers play a marginal role in the success of technology-based instruction. Since understanding the optimal interplay between instructors and software may allow to sustainably boost the cost-effectiveness of CAL, further research along these lines may prove valuable.

The documented educational value of a broadly available and offline-compatible learning software has direct policy implications: Most importantly, it highlights that CAL approaches may not only be an *effective* but also a *scalable* option for governments and NGOs operating in settings with low teacher prepardeness and poor internet coverage. When teachers are struggling with the concepts they have to teach, learning software can be an important remedy allowing them to improve the quality of their teaching. Another approach would be to invest in the skills of teachers, for instance by offering professional development programs: Teachers may not make much of a difference when they do not master what their students are supposed to learn, but vast empirical evidence from developed countries suggests that they can matter a great deal when they are well prepared and adequately qualified (Rockoff, 2004; Chetty et al., 2014). Hence, gaining a better understanding of how teachers' preparedness, and particularly their content knowledge, can be improved is likely to yield large social returns.

References

Adair, John G. 1984. The hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology* 69 (2):334–345.

Adair, John G., Donald Sharpe, and Cam-Loi Huynh. 1989. Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research* 59 (2):215–228.

Angrist, Joshua and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Attanasio, Orazio, Camila Fernández, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 2014. Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: Cluster randomized controlled trial. *BMJ* 349:1–12.

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Özler. 2018. Optimal design of experiments in the presence of interference. *The Review of Economics and Statistics* 100 (5):844–860.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics* 122 (3):1235–1264.

Banerjee, Abhijit V. and Esther Duflo. 2011. *Poor economics*. London: Penguin Books.

Bau, Natalie and Jishnu Das. 2020. Teacher value-added in a low-income country. *American Economic Journal: Economic Policy* 12 (1):62–96.

Beatty, Amanda, Emilie Berkhout, Luhur Bima, Thomas Coen, Menno Pradhan, and Daniel Suryadarma. 2018. Indonesia got schooled: 15 years of rising enrolment and flat learning profiles. RISE Working Paper No. 18/026.

Bettinger, Eric, Robert Fairlie, Anastasia Kapuza, Elena Kardanova, Prashant Loyalka, and Andrey Zakharov. 2020. Does EdTech substitute for traditional learning? Experimental estimates of the education production function. NBER Working Paper No. 26967, National Bureau of Economic Research, Cambridge, MA.

Beuermann, Diether W., Julian Cristia, Santiago Cueto, Ofer Malamud, and Yyannu Cruz-Aguayo. 2015. One laptop per child at home: Short-term impacts from a randomized experiment in Peru. *American Economic Journal: Applied Economics* 7 (2):53–80.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane. 2017a. Enrollment without learning: Teacher effort, knowledge and skill in primary schools in Africa. *Journal of Economic Perspectives* 31 (4):185–204.

Bold, Tessa, Deon Filmer, Ezequiel Molina, and Jakob Svensson. 2019. The lost human capital: Teacher knowledge and student achievement in Africa. World Bank Policy Research Working Paper No. 8849, World Bank Group.

Brunetti, Aymo, Konstantin Büchel, Martina Jakob, Ben Jann, Christoph Kühnhanss, and Daniel Steffen. 2020. Teacher content knowledge in developing countries: Evidence from a math assessment in El Salvador. Working Paper No. 2005, Department of Economics, University of Bern.

Carrillo, Paul E., Mercedes Onofa, and Juan Ponce. 2010. Information technology and student achievement: Evidence from a randomized experiment in Ecuador. IDB Working Paper Series No. 223, Inter-American Development Bank.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. Missing in action: Teacher and health worker absence in developing countries. *Journal of Economic Perspectives* 20 (1):91–116.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104 (9):2633–2679.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. 2017. Technology and child development: Evidence from the one laptop per child program. *American Economic Journal: Applied Economics* 9 (3):295–320.

de Ayala, R.J. 2009. *The theory and practice of item response theory.* New York: Guilford Press.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. 2014. Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education. In *Education policy in developing countries*, ed. Paul Glewwe. Chicago and London: University of Chicago Press, 285–338.

Digestyc, Direccion General de Estadistica y Censos El Salvador. 2018. Encuesta de hogares de direccion general de estadistica y censos 2017 (EHPM). Online avaliable: `www.digestyc.gob.sv`.

Escueta, Maya, Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quant. 2020. Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature* forthcoming.

Glennerster, Rachel and Kudzai Takavarasha. 2013. *Running randomized evaluations: A practical guide.* Princeton: Princeton University Press.

Glewwe, Paul and Karthik Muralidharan. 2016. Improving education outcomes in developing countries: Evidence, knowledge gaps and policy implications. In *Handbook of the economics of education*, eds. Eric Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: Elsevier, 653–743.

Heß, Simon. 2017. Randomization inference with Stata: A guide and software. *The Stata Journal* 17 (3):630–651.

Krueger, Alan B. 1999. Experimental estimates of education production functions. *The Quarterly Journal of Economics* 114 (2):497–532.

Lai, Fang, Renfu Luo, Lixiu Zhang, Xinzhe Huang, and Scott Rozelle. 2015. Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education Review* 47 (1):34–48.

Levitt, Steven D. and John A. List. 2011. Was there really a hawthorne effect at the hawthorne plant? An analysis of the original illumination experiments. *American Economic Journal: Applied Economics* 3 (1):224–238.

Linden, Leigh L. 2008. Complement or substitute? The effect of technology on student achievement in India. infoDev Working Paper No. 17.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. *The Quarterly Journal of Economics* 134 (3):1627–1673.

Mbiti, Isaac M. 2016. The need of accountability in education in developing countries. *Journal of Economic Perspectives* 30 (3):109–132.

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

Miguel, Edward and Michael Kremer. 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72 (1):159–217.

MINED, Ministerio de la Educacion de El Salvador. 2013. Elementos para el desarrollo del modelo pedagogico del sistema educativo nacional – escuela inclusiva de tiempo pleno. Online avaliable: `https://www.mined.gob.sv/jdownloads/Institucional/modelopedagogico.pdf`.

———. 2018. Informede resultados: PAES 2018. Online avaliable: `https://www.mined.gob.sv`.

Mo, Di, Linxiu Zhang, Jiafu Wang, Weiming Huang, Yao Shi, Matthew Boswell, and Scott Rozelle. 2015. Persistence of learning gains from computer assisted learning: Experimental evidence from China. *Journal of Computer Assisted Learning* 31:562–581.

Morgan, Kari Lock and Donald B. Rubin. 2012. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics* 40 (2):1263–1282.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review* 109 (4):1426–1460.

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

PAL, People's Action for Learning Newtork. 2020. International common assessment of numeracy. Online avaliable, URL: `https://palnetwork.org/ican/`.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review Papers and Proceedings* 94 (2):247–252.

Sacerdote, Bruce. 2011. Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the economics of education*, eds. Eric Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: Elsevier, 249–277.

Silm, Gerli, Margus Pedaste, and Karin Täht. 2020. The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review* 31:1–22.

Unesco, United Nations Educational, Scientific and Cultural Organization. 2019. UNESCO institute for statistics database. Online available: `http://data.uis.unesco.org/`.

The Economist. 2017. Technology is transforming what happens when a child goes to school. 22 July, Briefing Section.

World Bank. 2018. *World development report 2018: Learning to realize education's promise*. Washington D.C.: World Bank.

Yang, Yihua, Linxiu Zhang, Junxia Zeng, Xiaopeng Pang, Fang Lai, and Scott Rozelle. 2013. Computers and the academic performance of elementary school-aged girls in China's poor communities. *Computers & Education* 60 (1):335–346.

Young, Alwyn. 2019. Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics* 134 (2):557–598.