UNIVERSITÄT BERN

Faculty of Business, Economics and Social Sciences

Department of Economics

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

Konstantin Büchel, Martina Jakob, Christoph Kühnhanss, Daniel Steffen, Aymo Brunetti

20-06

April 2020

DISCUSSION PAPERS

Schanzeneckstrasse 1
CH-3012 Bern, Switzerland
http://www.vwi.unibe.ch

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

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

[a]Department of Economics, University of Bern
[b]Institute of Sociology, University of Bern
[c]Consciente – Unterstützungsverein für El Salvador, Bern

April 7, 2020

Abstract

This study provides novel evidence on the relative effectiveness of computer-assisted learning (CAL) software and traditional teaching. Based on a randomized controlled trial in Salvadoran primary schools, we evaluate three interventions that aim to improve learning outcomes in mathematics: *(i)* teacher-led classes, *(ii)* CAL classes monitored by a technical supervisor, and *(iii)* CAL classes instructed by a teacher. As all three interventions involve the same amount of additional mathematics lessons, we can directly compare the productivity of the three teaching methods. CAL lessons lead to larger improvements in students' mathematics skills than traditional teacher-centered classes. In addition, teachers add little to the effectiveness of learning software. Overall, our results highlight the value of CAL approaches in an environment with poorly qualified teachers.

JEL classification: C93, I21, J24, O15.

Keywords: computer-assisted learning, productivity in education, primary education, teacher content knowledge.

*We are grateful to Malin Frey, Amélie Speiser and David Burgherr 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 (ETH Zurich), 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, and the SSES Annual Congress in Geneva. This study would not have been possible without the generous *Impact Award 2017* prize money awarded by the SDC and NADEL (ETHZ), and co-funding by the chairs of Ben Jann (Insitute of Sociology, University of Bern) and Aymo Brunetti (Department of Economics, University of Bern). We received Institutional Review Board (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`.

Corresponding author: Konstantin Büchel, Postdoctoral Researcher at the Department of Economics, University of Bern; *Email:* konstantin.buechel@vwi.unibe.ch, *Phone:* +41 (0)31 631 49 97.

1 Introduction

While net primary school enrollment rates in low-income countries climbed from 56% in 2000 to 81% in 2019, the developing world still trails behind in terms of learning outcomes. 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, 2014), large and heterogeneous classes (Mbiti, 2016; Glewwe and Muralidharan, 2016), and low levels of effort among poorly trained teachers (Chaudhury et al., 2006; Bold et al., 2017a). A promising option to overcome these three barriers is to make greater use of computer-assisted learning (CAL). 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 if the software is used to personalize instructions.[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 is unclear whether learning gains are 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 is a substitute for certified teachers or if it is a complement to them. Finally, previous research has mostly evaluated specifically customized software which is 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 different treatments: The first treatment comprises additional math lessons instructed by a teacher (henceforth labeled as TEACHER). The second and third tre-

[1] Experimental studies on CAL interventions in low- and middle-income countries include Banerjee et al. (2007, math in Indian primary schools), Carrillo, Onofa and Ponce (2011, language and math in Ecuadorian primary schools), Yang et al. (2013, language and math in Chinese primary schools), Mo et al. (2015, math in Chinese primary schools), Lai et al. (2015, language and math in Chinese primary schools) and Muralidharan, Singh and Ganimian (2019, language and math with Indian secondary school pupils). They consistently report positive intent-to-treat estimates on learning outcomes that range between 0.1σ and 0.4σ.

[2] To our knowledge, the only study that evaluates the effectiveness of CAL lessons as a substitute to regular teaching 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 sample only covers NGO-run schools with well trained staff and innovative teaching methods. While it is unclear whether these findings translate to the challenging contexts of public education systems in developing countries, they still raise doubts about the inherent benefits of technology-based instruction.

atments are additional math lessons based on CAL software; one group of classes is monitored by technical supervisors (CAL + SUPERVISOR), while the other group is taught by teachers (CAL + TEACHER). Teachers had to be officially qualified to teach math in primary schools, whereas supervisors were laymen instructed to provide no content-related help to students. CAL lessons were taught using an offline application of the "Khan Academy" platform, and the three treatment arms were implemented by the Swiss-Salvadoran NGO *Consciente*.

We conducted the experiment between February and October 2018 with a sample of 198 primary school classes spanning grades 3 to 6 in the rural district of Morazán, El Salvador. 29 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 within-school control group (40 classes). In 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.

Main Findings. Our analysis establishes four key 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=0.00) when overseen by a supervisor and by 0.24σ (p-value=0.00) when instructed by teachers. These are intent-to-treat estimates reflecting an average attendance rate of about 59%. Using the treatment assignment as instrumental variable for attendance, we estimate that participating in all 46 additional CAL lessons (each lasting 90 minutes) translates to average learning gains of 0.36σ (p-value=0.00) and 0.38σ (p-value=0.00), respectively. This is equivalent to the average increase in math abilities over 1.2 school years.

Second, additional CAL lessons seem to have been 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=0.01). Hence, students assigned to CAL + TEACHER outperformed students assigned to TEACHER by 0.09σ (p-value=0.10); when analyzing percentage scores instead of standardized IRT-scores the according p-value decreases to 0.06. The CAL treatment overseen by technical supervisors (CAL + SUPERVISOR) was also slightly more successful in raising student learning than traditional teaching, even though this difference clearly falls short of statistical significance (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 classes instructed by teachers and to 0.09σ (p-value=0.12) for the CAL lessons monitored by supervisors.

Third, we present multifaceted evidence that points to a rather low productivity of teachers. The difference in learning gains between within-program school control classes and those classes receiving additional teacher-centered math lessons was close to zero and statistically insignificant (p-value=0.78). Similarly, teachers did not provide much "value added" to the learning software: the estimated impact for CAL lessons instructed by teachers is slightly higher than for CAL lessons conducted by supervisors but the difference is negligible and statistically insignificant (p-value=0.65). Moreover, the productivity of teachers dropped as the complexity of concepts increased: The impact of additional math lessons instructed by teachers decreased sharply with both the grade level and the baseline achievement of their students, while the effect of the CAL-based lessons was largely insensitive to students' grades and initial ability levels. To gain a better understanding of the mechanisms behind these findings, we conducted a comprehensive teacher math assessment covering the primary school curriculum of El Salvador. This assessment documents very poor content knowledge among the teachers hired by the NGO. Furthermore, regular math teachers in local primary schools are even less proficient in their subject. Potential productivity gains resulting from an introduction of CAL to regular classes may thus be larger than suggested by our estimates, since the NGO's contract teachers had better content knowledge and employed more modern pedagogical techniques than regular math teachers.

Finally, we document substantial treatment externalities. At endline, students in within-program school control classes outperformed 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 social learning. At the same time, the data rejects hypotheses operating via direct exposure of control students to the treatments (i.e. non-compliance) or behavioral adjustments in response to the experimental design.

Contributions. 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. To our knowledge, this is the first well-identified study assessing the value-added of CAL in a public school setting of a developing country. 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. While CAL has been regularly praised in terms of its individualized and interactive pedagogy (e.g. Banerjee et al., 2007; Muralidharan, Singh and Ganimian, 2019), our findings highlight that it may also be a promising approach to mitigate the adverse effects of teachers' inadequate content knowledge and pedagogical knowledge, that has been recently documented for several developing countries (e.g. Bold et al., 2017a).

Second, we also present the first experimental test of the complementarities between teachers and learning software. In our setting, teachers seem to play a marginal role in the success of technology-based instruction, with CAL lessons being almost equally effective when provided by a supervisor rather than a certified teacher. Thus, teachers and learning software are likely substitutes and not complements. 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., 2015). 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 credible 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 freely available and features extensive math contents in more than 30 languages.[3] Since the employed software is arguably one of the most important features of a CAL intervention, our findings bear direct policy relevance for educational decision-makers around the globe that are looking for a readily available learning software suitable in non-English speaking countries.[4]

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 get to attend primary school, access becomes more selective at later stages of an educational career with secondary and tertiary enrollment standing at 67% and 28%, respectively.[5]

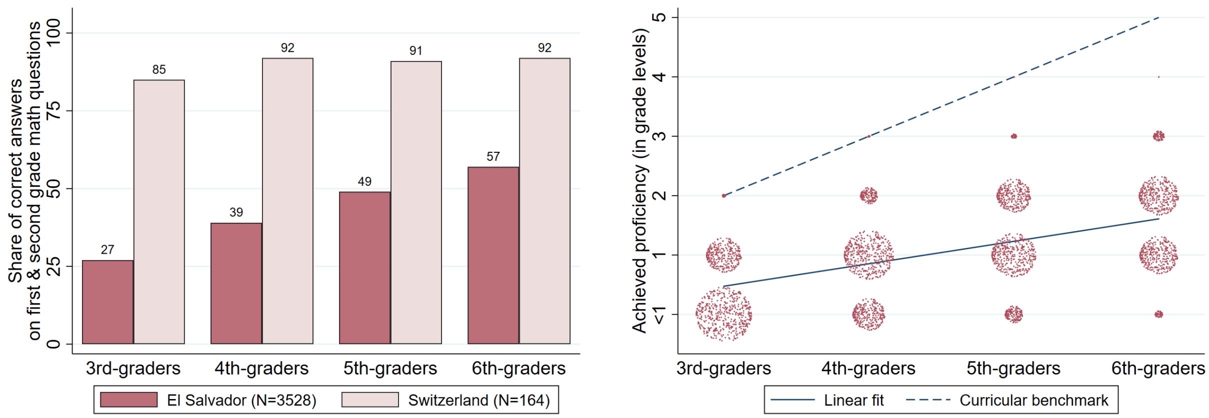
The department of Morazán is a poor and rural region in the northeast of the country with roughly 200,000 inhabitants. An average person in Morazán lives on 3.80 USD per day and, according to national definitions, almost 50% of the households face multifaceted poverty. With an illiteracy rate of more than 20%, Morazán ranks second-last among all Salvadorian departments in terms of educational attainment (Digestyc, 2018).

Our math assessments with 3,528 third to sixth graders conducted in February 2018 further reveal that primary school children 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.

[3] The full version is available in 16 languages including Spanish, English, Chinese, French and Portuguese. A subset of content is available in another about 20 languages including Russian, Hindi and Swahili. For further information see the Khan Academy website `https://www.khanacademy.org/` (last visit: 01.12.2019).

[4] Another off-the-shelf learning software that has been successfully evaluated is Mindspark (see Muralidharan, Singh and Ganimian, 2019), which operates in English and Hindi for math and language training. Other studies evaluate customized software that is not readily available, for instance Yang et al. (2013) or Lai et al. (2015).

[5] 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.

Several challenges that plague Morazán's schooling system can help to explain its low productivity. For instance, our monitoring data from school visits reveal high rates of teacher absenteeism so that, on average, 25% of regular lessons are canceled. Low teacher motivation mixes with outdated pedagogical techniques that basically follow the logic of "copy, learn by heart, 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 similar issues to those reported for other low- and middle income countries.[6]

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

[6]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, Singh and Ganimian (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 3 grade levels for primary schools. Moreover, Pritchett and Beatty (2014) 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.

morning or afternoon lessons throughout El Salvador, the new SI-EITP policy[7] 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 criminal gangs. Within the scope of this countrywide program, the Ministry of Education seeks to cooperate with NGOs in order to collectively promote an open and flexible curriculum. While all schools received official 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 3 to 6. The program features three intervention arms, that offer two additional lessons of 90 minutes per week and almost double the beneficiaries' number of math lessons 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 computer-assisted learning software; one group of classes is taught by teachers, while the other group is instructed by supervisors.

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 *K-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 working plans covering different content units was prepared. Based on a placement test, children received a plan that was viewed as accurate 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 grades.

A similar methodology was used for the first intervention arm that features more traditional *math lessons instructed by a teacher*. According to their initial math skills, children were arranged in two different 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 necessary. 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, concentration 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 or at the end of each session. While supervisors were instructed to use animation and

[7] SI-EITP stands for *Sistema Integrado de Escuelas Inclusivas de Tiempo Pleno*, which translates to *Integrated System of Inclusive Full Time Schools.*

concentration games, teachers were additionally introduced to a series of math games.

The contracted *teachers* were required to be officially qualified to instruct grade 3 to 6 children 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 assuming four classes.[8] 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 meetings and unannounced classroom visits during the implementation phase.

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 have minimal IT skills and some experience in dealing with children, while teaching credentials and a specific educational degree were not among the selection criteria. 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.[9] 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 3 to 6, and therefore can participate with at least (*i*) one CAL+TEACHER, (*ii*) one CAL+SUPERVISOR, (*iii*) one TEACHER, and (*iv*) one control class;

[8] This corresponds to 8 × 90 minutes of teaching per week, or – including preparatory work – to a 60% job. A smaller group of teachers only assumed two classes (i.e 4 × 90 minutes of teaching per week, or approx. a 30% job).

[9] The school year in El Salvador starts in mid-January and ends in November.

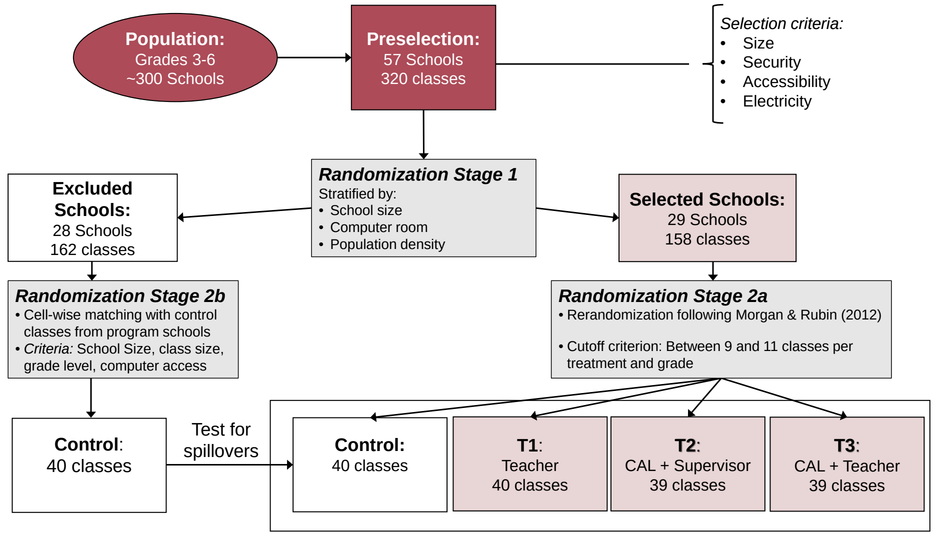


Figure 2: Sampling and randomization scheme.

- ***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 that are hardly accessible by car 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 6400 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. We account for the re-randomization 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, p. 572), who finds no appreciable change in rejection rates beyond this threshold. To implement the randomization tests we rely on Stata's RITEST-package developed by Hess (2017).

As prominently discussed in Miguel and Kremer (2004), interventions can have spill-over 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 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 on any observable dimension at baseline, Table 1 shows a sizeable increase in the absence rates between baseline and 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.[10] Moreover, compliance with the 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 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

[10]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	Within 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)	
Std. 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 and absence rates during assessments (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	0.73	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)				Treatment 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.

official textbooks as well as various international sources of student assessments; the Appendix Section B 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.2 shows that virtually all of the items (except one for fifth graders in the endline assessment) were at least once answered correctly or incorrectly. Likewise, Table B.1 documents that only about 0.5 percent 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 percent (3rd graders) to 40 percent (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 techniques from psychology called *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 base- and 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 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 school. While the government gathers thematically broad data 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 utilizable information on the class structure, enumerators had to call each school during the first weeks of January, 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 3 to grade 6 classes (school size), information on the presence of gangs (security at school), accessibility measures based on Google-Map 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. They covered both regular lessons as well as program lessons and 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 within-program school control classes (i.e. β_{CX}) using

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

where Y_{ics}^{EL} is the endline math score of student i in class c and school s; Math scores are either measured as percentage of correct answers or the IRT-score normalized to $\mu = 0$ and $\sigma = 1$ 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 conducted by a supervisor, $T3$ equals one for those assigned to extra CAL lessons conducted by a teacher, and CX equals one for those assigned to within-program school control classes that are potentially subject to externalities. Our control variables include Y_{ics}^{BL} that stands for the baseline math score, X_{ics} representing a set of student-level control variables (i.e. age standardized by average grade age, gender, household size and household assets), and V_{cs} 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_{1ics} represents the error term.

The upper 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 lower panel of Table 2 presents estimates for the pairwise differences between the three treatment groups in program schools. The lower panel reports 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). In the upper panel, however, p-values are based on traditional clustered standard errors, since the assignment to program schools and pure control schools did not involve re-randomization.[11]

Students who were assigned to one of the treatments perform significantly better in the end-

[11] Moreover, we cannot properly apply randomization inference to the upper panel due to missing information on ability levels of non-selected classes from pure control schools. As we show in Appendix A.3, randomization inference in the upper panel 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 clustered standard errors. While this is clearly too conservative, our main conclusion are not altered when we apply randomization inference to the upper panel (see Table A.2). 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 lower panel, as shown in Table A.3, 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)	(5)	(6)
T1: Lessons with Teacher	2.904***	2.643**	0.165***	0.152**
	(0.005)	(0.012)	(0.006)	(0.013)
T2: CAL-Lessons with Supervisor	4.095***	3.869***	0.226***	0.214***
	(0.000)	(0.000)	(0.000)	(0.000)
T3: CAL-Lessons with Teacher	4.554***	4.328***	0.250***	0.238***
	(0.000)	(0.000)	(0.000)	(0.000)
CX: Control Classes for Externalities	2.595**	2.407**	0.147**	0.137**
	(0.011)	(0.017)	(0.013)	(0.020)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	1.191	1.226	0.061	0.063
p-value (β_{T4}=0)	(0.214)	(0.194)	(0.268)	(0.244)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	1.650*	1.686*	0.084	0.086
p-value (β_{T5}=0)	(0.069)	(0.059)	(0.117)	(0.102)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.459	0.460	0.024	0.023
p-value (β_{T6}=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: In the upper panel (coef. $\beta_{T1} - \beta_{CX}$), p-values are based on traditional clustered standard errors. In the lower panel (coef. $\beta_{T4} - \beta_{T6}$), p-values 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<0.10, ** p<0.05, *** p<0.01.

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

Finally, we test whether the observed gaps in the endline performance of students assigned to one of the three treatments (defined as β_{T4}, β_{T5}, and β_{T6}) are statistically different from zero. While we find that the two CAL treatments outperform additional math classes, only the difference 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 measurable learning gains in a Latin American context using off-the-shelf learning software. And the impact of additional CAL lessons is considerable: Expressing the impact 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). On the other hand, traditional math classes conducted by teachers are relatively ineffective compared to additional math lessons with CAL-software: In comparison to the within program school control classes, boosting the supply of conventional math lessons by roughly 80% delivered no measurable impact. Importantly, 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 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 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 about 10% to 20% compared to the overall impact reported in Table 2. The NSEA math score of students assigned to CAL classes with teachers is 4.9 percentage points or 0.26σ higher than the score of students assigned to pure control classes; again this effect is 10% to 15% larger compared to estimates based on all questions. Since the impact on the NSEA math score remains about the same for students receiving additional math classes instructed by teachers, the gap between CAL and conventional teaching widens.

When we compare the learning gains attributed to CAL with the gains attributed to the additional math classes without software the differences range between 1.7 and 2.1 percentage points or 0.092σ and 0.115σ. The corresponding p-values lie in between 0.046 and 0.055 for the CAL classes

Table 3: ITT-Estimates on the effects of the interventions on children's *NSEA*-scores

	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teacher	3.174***	2.849***	0.166***	0.146**
	(0.002)	(0.006)	(0.006)	(0.013)
T2: CAL-Lessons with Supervisor	4.907***	4.581***	0.258***	0.238***
	(0.000)	(0.000)	(0.000)	(0.000)
T3: CAL-Lessons with Teacher	5.225***	4.895***	0.279***	0.259***
	(0.000)	(0.000)	(0.000)	(0.000)
CX: Control Classes for Externalities	2.711***	2.463**	0.145**	0.130**
	(0.008)	(0.012)	(0.013)	(0.020)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	1.733	1.732*	0.092	0.091
p-value (β_{T4}=0)	(0.103)	(0.093)	(0.129)	(0.115)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	2.051**	2.047**	0.113*	0.112*
p-value (β_{T5}=0)	(0.046)	(0.047)	(0.051)	(0.055)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.318	0.315	0.021	0.021
p-value (β_{6}=0)	(0.750)	(0.752)	(0.706)	(0.714)
Adjusted R^2	0.62	0.63	0.65	0.65
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: In the upper panel (coef. $\beta_{T1} - \beta_{CX}$), p-values are based on traditional clustered standard errors. In the lower panel (coef. $\beta_{T4} - \beta_{T6}$), p-values 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<0.10, ** p<0.05, *** p<0.01.

with teachers and between 0.093 and 0.129 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.

Table 4 shows the results that are based on GEOM- and DSP-items. Focusing on these topics mitigates 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

Table 4: ITT-Estimates on the effects of the interventions on children's *GEOM & DSP*-scores

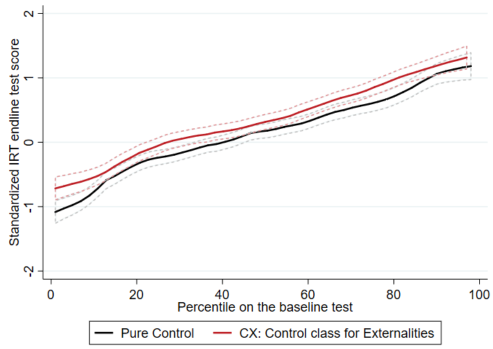
	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teacher	2.433*	2.132*	0.155**	0.140*
	(0.055)	(0.093)	(0.035)	(0.057)
T2: CAL-Lessons with Supervisor	3.207***	3.014**	0.196***	0.187***
	(0.009)	(0.014)	(0.006)	(0.009)
T3: CAL-Lessons with Teacher	3.646***	3.472***	0.201***	0.193**
	(0.006)	(0.008)	(0.008)	(0.010)
CX: Control Classes for Externalities	2.773**	2.561**	0.159**	0.149**
	(0.032)	(0.048)	(0.036)	(0.050)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	0.775	0.882	0.041	0.047
p-value (β_{T4}=0)	(0.498)	(0.432)	(0.543)	(0.464)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	1.213	1.340	0.046	0.053
p-value (β_{T5}=0)	(0.279)	(0.221)	(0.481)	(0.412)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.438	0.458	0.005	0.006
p-value (β_6=0)	(0.692)	(0.669)	(0.934)	(0.926)
Adjusted R^2	0.46	0.47	0.49	0.50
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: In the upper panel (coef. $\beta_{T1} - \beta_{CX}$), p-values are based on traditional clustered standard errors. In the lower panel (coef. $\beta_{T4} - \beta_{T6}$), p-values 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<0.10, ** p<0.05, *** p<0.01.

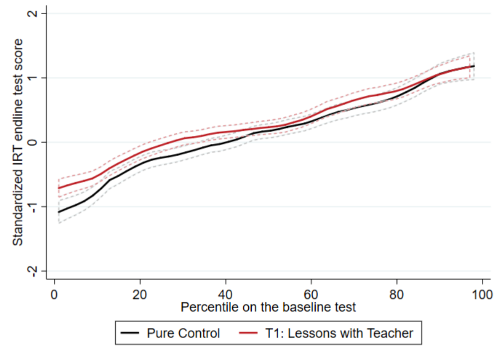
the lower-bound effects net of any spillovers are consistently significant for CAL + TEACHER and just at the edge of the 0.1-threshold for CAL + SUPERVISOR, the measured differences seem primarily driven by the pronounced improvements in the domains of number sense and elementary arithmetic. The intervention was less successful in shifting abilities to solve questions on geometry, measurement, data and statistics: the point estimates decrease by about 30 percent, and the p-values clearly exceed the 0.1-threshold for statistical significance. This sub-analysis also confirms the strikingly low productivity of certified teachers: Whether we compare the performance across items on basic arithmetic or items on geometry and data analysis, classes receiving additional math lessons conducted by teachers do not perform better than control classes subject to externalities.

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

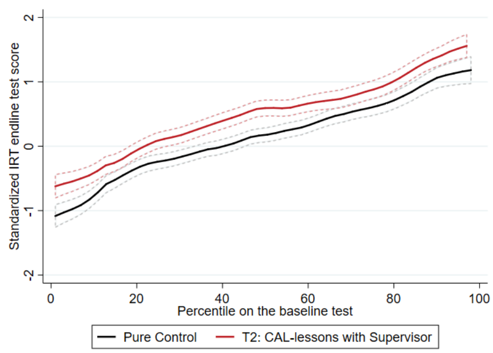
We continue the heterogeneity analysis by discussing Figure 3, which plots kernel-weighted locally-smoothed means of the endline test score at each percentile of the baseline test score by treatment



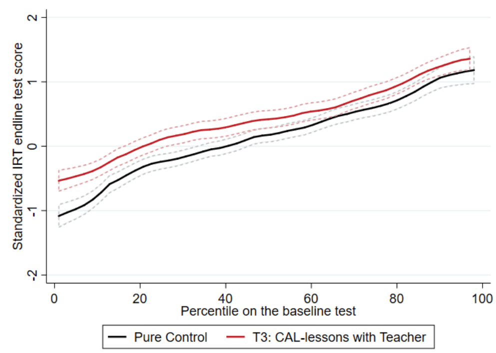
(a) Spillover vs. pure control classes



(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.

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 effect and baseline achievement more closely. Similarly, we further investigate whether the reported effects vary by grade level or class size. To do so, we estimate

$$\begin{aligned}Y_{ics}^{EL} = \alpha + \beta_{T1}T1_{cs} + \beta_{T2}T2_{cs} + \beta_{T3}T3_{cs} + \beta_{CX}CX_{cs} \\ +\theta_1(T1_{cs} \times Var_{ics}) + \theta_2(T2_{cs} \times Var_{ics}) \\ +\theta_3(T3_{cs} \times Var_{ics}) + \theta_{CX}(CX_{cs} \times Var_{ics}) \\ +\delta Y_{ics}^{BL} + X_{ics}^{'}\gamma + V_{cs}^{'}\lambda + \phi_k + \epsilon_{2ics}\end{aligned} \tag{2}$$

where $(T_{cs} \times Var_{ics})$ 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).

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 5). 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) and T3 (i.e. CAL + TEACHER), but the p-values do not reach the 10%-threshold. 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.

Table 5: Effect heterogeneity along baseline ability, grade level and class size.

Treatment indicators interacted with: *Dependent variable: Std. IRT-Score*	Baseline Math Score (1)	Grade Level (2)	Class Size (log) (3)
T1: Lessons with Teacher × Var.	−0.105***	−0.140***	−0.436***
	(0.004)	(0.000)	(0.005)
T2: CAL-Lessons with Supervisor × Var.	−0.014	−0.051	−0.110
	(0.739)	(0.252)	(0.433)
T3: CAL-Lessons with Teacher × Var.	−0.038	−0.058	−0.270*
	(0.285)	(0.184)	(0.052)
CX: Classes exposed to Externalities × Var.	−0.005	−0.022	−0.118
	(0.910)	(0.688)	(0.483)
Adjusted R^2	0.70	0.70	0.70
Observations	3197	3197	3197
Baseline Score	Yes	Yes	Yes
Individual & Classroom Controls	Yes	Yes	Yes
Stratum & Grade Level FE	Yes	Yes	Yes

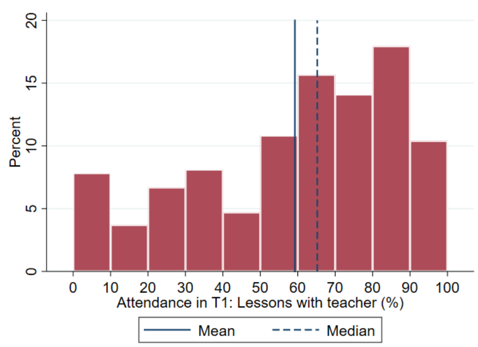
Notes: p-values are based on class-level clustered standard errors and are shown in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

A similar pattern emerges when we study effect heterogeneity by grade level of the participating students (see column 2 in Table 5). The effects of the CAL treatments do not significantly vary with the grade level of students, but we find that additional math lessons taught by a teacher are least effective in higher grades. This corroborates the finding that without the help of learning software, teachers in Morazán seem to be least effective when explaining more complex concepts.

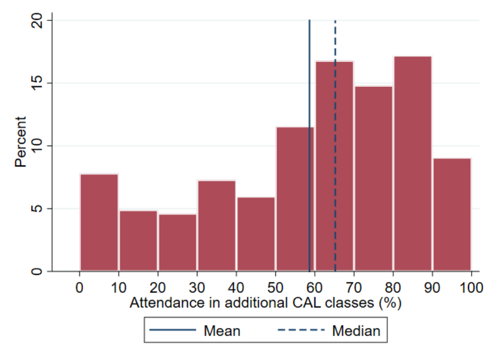
Finally, we find that large class sizes reduce the effectiveness of teachers (see column 3 in Table 5), no matter whether they use CAL software or not. Interestingly, this pattern does not emerge for CAL classes instructed by a supervisors. This seems plausible, since supervisors were directed to refrain 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 ($\hat{\theta}_1$=0.436, p-value=0.005) is more sensitive to class size than the effect of CAL-lessons instructed by teachers ($\hat{\theta}_3$=0.270, p-value=0.052). Overall, this confirms the notion that computer-based learning can mitigate the problems related to large class sizes (e.g. Banerjee and Duflo, 2011; Muralidharan, Singh and Ganimian, 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 therefore present data on the overall compliance, examine the correlation between individual attendance rates and individual endline scores, and finally discuss instrumental variable estimates for the impact of the three interventions assuming full attendance.



(a) T1: Additional math lessons with teachers



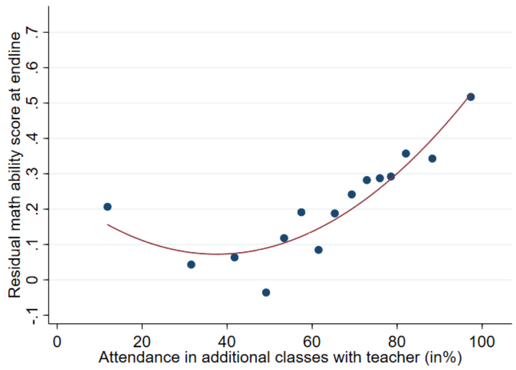
(b) T2 & T3: Additional CAL-lessons

Figure 4: Attendance of students in additional math lessons.

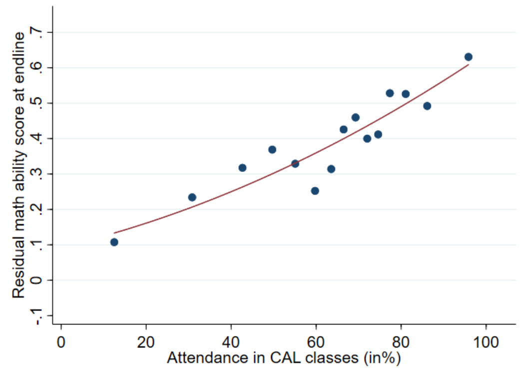
Figure 4 plots the distribution in attendance rates across all eligible students. With an average attendance rate of 59%, participation of students was a weak spot of the program. Attendance rates slightly varied across the three treatments, although the differences are statistically insignificant: Additional 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%).

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. 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 those 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.[12]

[12] 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: $\hat{\gamma}_{T1}$=0.42 (t-value=4.6); $\hat{\gamma}_{T2}$=0.53 (t-value=3.8); $\hat{\gamma}_{T3}$=0.58 (t-value=3.8). Including a quadratic term we get: $\hat{\gamma}^1_{T1}=-0.67$ (t-value$=-2.3$), $\hat{\gamma}^2_{T1}$=1.04 (t-value=3.5); $\hat{\gamma}^1_{T2}$=0.55 (t-value=1.1), $\hat{\gamma}^2_{T2}=-0.03$ (t-value$=-0.1$); $\hat{\gamma}^1_{T3}=-0.40$ (t-value$=-0.9$), $\hat{\gamma}^2_{T3}$=0.96 (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.

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

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

where $Att_i^{T=t}$ is student's i attendance rate in treatment t and takes values between 0 and 1. 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. $\widehat{Att}_{ics}^{T=t}$, and estimate

$$Y_{ics}^{EL} = \alpha + \beta_1 \widehat{Att}_{ics}^{T=1} + \beta_2 \widehat{Att}_{ics}^{T=2} + \beta_3 \widehat{Att}_{ics}^{T=3} + \delta Y_{ics}^{BL} + X_{ics}^{'}\gamma + V_{cs}^{'}\lambda + \phi_k + \epsilon_{4ics}. \tag{4}$$

In order to interpret $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ 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, Singh and Ganimian, 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 (both CAL treatments) or decreasing (TEACHER) in initial ability (see section 4.2.2), even though attendance rates are positively correlated with baseline scores. Attending an additional math lesson thus had a stronger effect on low ability than high ability students. Hence, the IV-estimates might undervalue 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

Table 6: IV-Estimates: Program effects with full participation

	Percent Correct		Std. IRT-Scores	
	(1)	(2)	(3)	(4)
T1: Lessons with Teacher	5.204***	4.721***	0.283***	0.255**
	(0.001)	(0.005)	(0.004)	(0.010)
T2: CAL-lessons with Supervisor	7.113***	6.762***	0.377***	0.357***
	(0.000)	(0.000)	(0.000)	(0.000)
T3: CAL-lessons with Teacher	7.686***	7.275***	0.404***	0.379***
	(0.000)	(0.000)	(0.000)	(0.000)
Kleibergen-Paap F-statistic	225.85	204.47	225.48	203.96
Adjusted R^2	0.65	0.66	0.68	0.69
Observations	2570	2570	2570	2570
Baseline Score	Yes	Yes	Yes	Yes
Individual & Classroom Controls	No	Yes	No	Yes
Stratum & Grade Level FE	Yes	Yes	Yes	Yes

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

additional math lessons. Again this would lead to a downward bias in the reported IV-estimates.

Table 6 presents the IV-estimates, that 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 0.36σ to 0.40σ, which is about equivalent to the average student's progress in 1.2 school years.[13] This is comparable in magnitude to effects of technology-aided instructions found in India, where Muralidharan, Singh and Ganimian (2019) report average learning gains in math of 0.6 standard deviations for a 90 days attendance at CAL learning centers.

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 different sources allows for a discussion of what may (or may not) explain these externalities. In the following we

[13] We refrain from presenting F-tests that formally test whether the difference between the three interventions are statistically significant because our re-randomization scheme for the within school assignment of treatments would require randomization inference, which we cannot implement in the IV-setting.

distinguish between three broad explanations: *(i)* direct exposure of control students, *(ii)* behavioral adjustments to the experimental design, and *(iii)* social learning among peers.

Direct Exposure. We begin with 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 official participants. Our monitoring data shows that compliance with this directive was high, as unauthorized participation was only recorded during 6 out of about 750 unannounced visits in NGO-run math 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 (about 1%) students in our estimation sample changed classes or schools during the program and excluding these students from the estimation models leaves the results virtually unchanged.

Control students in program classes may also have been exposed to the learning software in regular classes or at home. Again, our data suggests otherwise: The enumerators recorded computer usage in only 5 out of about 1,000 regular class visits. Similarly, computer usage at home is an unlikely candidate to account for treatment externalities: According to our socio-demographic survey, only 576 students (about 18%) live in a household that owns a computer with internet access and this asset class is not correlated with learning outcomes in the endline assessment.

Behavioral Adjustments to the Experimental Design. We now discuss the likelihood of behavioral adjustments of teachers and students to the experimental design, namely unintended incentives to improve performance at the school level or reactive behavior of the control group.

The presence of the NGO might have incentivized school staff 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. We first examine this reasoning by using class cancellation and attendance rates as proxies for the effort by school staff/teachers, and then continue by testing whether a more generous supply of computer hardware raised performance in control classes. Contrary to expectations, cancellation rates appear to be slightly higher in program schools than in control schools although the difference is not statistically significant (see columns 4 & 5 in Table 7).[14]

[14] The project could also have affected class cancellation rates directly, e.g. due to space limitations inducing the conduction of the additional lessons at the expense of regular classes. Furthermore, differences in recorded cancellation rates may (partly) be an artifact of the data collection process. To minimize transport expenses, we randomly selected entire *schools* rather than *classes* to be visited on a given day. Thus, enumerators were faced with slightly different settings in treatment and control schools: They had to record data from all classes on grades 3–6 in treatment schools, and only about one to two classes during visits to control schools. One could hypothesize that, in control schools, data collectors might have been inclined to wait patiently for the teacher to turn up (to be able to conduct the classroom observations), while, in treatment schools, they may have moved on to the next class.

Table 7: Externality channel (II): Motivation proxied with class attendance and cancellations.

Dependent variable:	Student Attendance (%)			Class Cancellations (%)		
	(1)	(2)	(3)	(4)	(5)	(6)
Program Schools	−0.311	−0.295	−0.978	6.884	6.534	8.215
	(0.889)	(0.893)	(0.648)	(0.237)	(0.264)	(0.140)
Adjusted R^2	0.07	0.06	0.00	0.08	0.08	0.08
Observations	198	198	80	198	198	80
Control Classes Only	No	No	Yes	No	No	Yes
Classroom Controls	No	Yes	Yes	No	Yes	Yes
Grade Level FE	Yes	Yes	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes

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

Similarly, student attendance rates do not point towards intensified efforts in program schools, as the estimated differences in columns (1) and (2) of Table 7 yield p-values larger than 0.8. Finally, we test whether a more generous furnishing of computer-labs by the NGO has pushed schools to better performances that is not necessarily reflected in attendance and cancellation rates. Consistent with the previous results, columns (3) and (4) in Table 8 show no relevant correlation between the number of NGO computers installed in a school and the endline performance of students in control classes.

The difference between control classes within an outside treatment schools could also be driven by a so-called *John Henry Effect*: a bias resulting from reactive behavior of the control group (e.g. Glennerster and Takavarasha, 2013). In our setting, such a bias could result either from student or from teacher behavior. As to the former, students in control classes might have worked harder to make up for their disadvantage. Similarly, teachers could have redirected resources and effort towards control classes to compensate them for their relative deprivation. For example, teachers may have given more weight to math relative to other subjects when attending control classes. If such behavior arises within treatment schools, but not in geographically (and thereby socially) separated schools, it could account for the observed treatment externalities. This mechanism has similar implications, but is distinguishable from those discussed in the previous paragraph. While the last 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 7, limiting the analysis to the control classes does not alter our conclusions: The difference in class cancellation rates between program school control classes and pure control classes is small and remains aloof from any conventional level of statistical significance. The same applies for students' attendance rates.

Social Learning among Peers. The treatment externalities may also stem from *social learning and peer effects*, as participating students could have shared their knowledge and motivation with

Table 8: Externality channel (I): Proxies for social learning and in-kind incentives

Dependent variable: Std. IRT Score	Treatment Intensity		Installed NGO computers	
CX-indicator interacted with:	All Treatments	CAL	Per Student	Total
	(1)	(2)	(3)	(4)
CX: Control Classes for Externalities	0.146**	0.135**	0.142**	0.146**
	(0.019)	(0.023)	(0.020)	(0.037)
CX: 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	No	No	No	No
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

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

their peers from other classes. Results in columns (1) and (2) of Table 8 suggest that this may have been the case: What explains part of the performance differential between within-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 (partly) passed on by the participants to their peers from non-treated classes. Another explanation for this pattern would be that hosting many CAL classes came about 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. However, as discussed above, columns (3) and (4) in Table 8 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 to their peers seems more plausible than behavioral adjustments in prospect of being donated new equipment. This finding is consistent with a broad literature of peer-effects that documents how the performance of each student affects achievements of their class-mates (see Sacerdote, 2011).

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 the monitoring 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 operating via school teachers' and students' attendance (a proxy for motivation), but the data consistently rejects this set of claims. Considering that social learning remains as the most plausible explanation for the observed treatment externalities further strengthens the case in favor of the CAL interventions.

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 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 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. Note, however, that these conclusions have to be interpreted with care: Not only is precision impaired by the statistical uncertainty of our impact estimates, but relative cost-effectiveness is also dependent on different contextual factors such as the local wage levels, the wage premium for certified teachers or the availability of affordable hardware.

5.3 The Role of Teacher Ability

Multifaceted evidence derived in our analysis points to a relatively low productivity of teachers. *First*, the difference in learning gains between within-program school control classes and classes receiving additional teacher-centered math lessons are close to zero and statistically insignificant (p-values around 0.7). Similarly, 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 slightly higher than for CAL lessons conducted by supervisors but statistically speaking they are not distinguishable (again the p-values are in the 0.7 range). *Second*, 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. *Third*, both CAL interventions (at least marginally) outperform the additional math lessons instructed by teachers: The point estimates of the CAL interventions are consistently larger, and the impact of neither CAL + TEACHER nor CAL + SUPERVISOR decreases with student baseline performance or grade level. Hence it appears that in our setting, learning software is more productive in teaching basic math than certified teachers, especially as the complexity

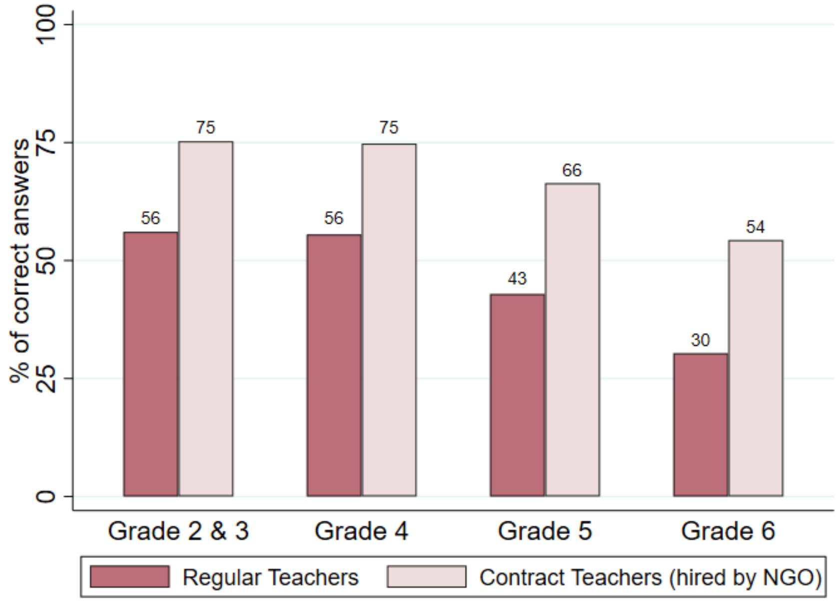


Figure 6: Math proficiency among regular teachers and teachers hired for additional math lessons. *Note:* The graph shows the share of correct answers on questions covering the official math curriculum of grades 2 & 3, grade 4, grade 5, and grade 6. This data was collected after the endline assessment 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. *Source:* Brunetti et al. (2020).

of the contents increases.

In order to analyze the root cause of the low productivity of teachers, we asked the instructors hired by the NGO to participate in an 90 minutes math assessment covering the primary school curriculum of grades 2 to grade 6. Moreover, we administered the same assessment to a representative sample of regular math teachers of grade 3 to grade 6 classes which allows us to learn how the contract teachers compare to the regular teaching staff (see Brunetti et al., 2020, for details on the assessment). Figure 6 illustrates the main insights from this assessment: primary school math teachers in the department of Morazán insufficiently master the contents 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 in recent World Bank contributions (see Bold et al., 2017b; World Bank, 2018). This direct evidence on the lack of content knowledge conforms with our finding on the teachers' low productivity in conveying math concepts, especially those concepts pertaining to higher grades.[15]

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

[15]Since the teachers performed considerably worse than expected, we also validated the questions by administering the identical test (translated to German) to 16 Swiss primary school teachers, who achieved a median score of 90% (i.e. 45 correct answers out of 50 questions).

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 alarmingly low 30% on items pertaining to the sixth grade curriculum.[16]

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 even too optimistic when aiming for a scale up with regular teachers, who have on average an even lower math proficiency than the (much younger) contract teachers hired by the implementing NGO. At the same time, these results highlight the value of learning software that 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, Singh and Ganimian, 2019). While our heterogeneity analysis corroborates this line of reasoning, we further show that computer-assisted learning can help to remedy shortcomings related to low teacher ability. Since teachers are considered to be the most pivotal input to the learning production function (e.g. World Bank, 2018, p. 80), these findings also raise the question, how teacher quality can be improved in an effective manner.

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 can not only produce substantial learning gains, but may also outperform traditional instruction. In our setting, this relative advantage seems to be driven by a mismatch between teacher preparation and the complexity of the concepts they have to teach: Under traditional teaching models, it seems questionable that children are able to master what their teachers fail to understand, while CAL allows them to make progress beyond their teachers' content knowledge. Overall, our findings point to an alarmingly low productivity of teachers. Not only is the effect of additional teacher-led instruction comparatively low (and might be partly if not completely attributable to treatment externalities), but poorly trained teachers also do little to improve the productivity of CAL lessons. In light of the fact that they do not master a large share of the contents they are required to teach, these results are hardly surprising.

Promoting the targeted use of computers may therefore be an attractive option for governments and NGOs operating in settings with low teacher quality. When teachers are struggling with

[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 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.

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. However, 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, Friedman and Rockoff, 2014). Hence, gaining a better understanding of how teachers' preparedness, and particularly their content knowledge, can be improved seems to be crucial for researchers as well as policy makers. Since hardly any rigorous evidence on this aspect is available (see Muralidharan, 2017; Bold et al., 2017a), we teamed up with the same implementing partner to examine in an ongoing study, whether computer-assisted learning software can help to effectively improve the content knowledge of teachers and therewith their productivity in the classroom (see Brunetti et al., 2019).

References

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, Aislinn Bohren, Craig McIntosh and Berk Özler. 2015. "Designing Experiments to Measure Spillover Effects." PIER Working Paper No. 15-021. URL: `www.ssrn.com/` (last access: 10.03.2020).

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

Banerjee, Abhijit, 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.

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, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian Stacy, Jakob Svensson and Waly Wane. 2017b. "What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa." World Bank Policy Research Working Paper No. 7956. URL: `www.worldbank.org/` (last access: 10.03.2020).

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. URL: `www.vwi.unibe.ch` (last access: 10.03.2020).

Brunetti, Aymo, Konstantin Büchel, Martina Jakob, Ben Jann, Christoph Kühnhanss and Daniel Steffen. 2019. "How Effective Are Computer-Based Teacher Training Programs? Evidence from a Randomized Controlled Trial in El Salvador." *AEA RCT Registry*: `https://doi.org/10.1257/rct.4092-2.0` (last access: 05.03.2020).

Carrillo, Paul, Mercedes Onofa and Juan Ponce. 2011. "Information Technology and Student Achievement: Evidence from a Randomized Experiment in Ecuador." IDB Working Paper Series No. 223. URL: `www.econstor.eu` (last access: 10.03.2020).

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan and 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 Friedman and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9):2633–2679.

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 pp. 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, URL: `www.digestyc.gob.sv` (last access: 25.07.2018).

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*, ed. Eric Hanushek, Stephen Machin and Ludger Woessmann. Amsterdam: Elsevier pp. 653–743.

Hess, Simon. 2017. "Randomization Inference with Stata: A Guide and Software." *The Stata Journal* 17(3):630–651.

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." *Ecomomics of Education Review* 47(1):34–48.

Linden, Leigh. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India." infoDev Working Paper No. 17. URL: `www.worldbank.org/` (last access: 10.03.2020).

Mbiti, Isaac. 2016. "The Need of Accountability in Education in Developing Countries." *Journal of Economic Perspectives* 30(3):109–132.

Mbiti, Isaac, Karthik Muralidharan, Maruicio 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.

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 available, URL: `https://www.mined.gob.sv/jdownloads/Institucional/modelopedagogico.pdf` (last access: 14.01.2018).

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 and Donald Rubin. 2012. "Rerandomization to Improve Covariate Balance in Experiments." *The Annals of Statistics* 40(2):1263–1282.

Muralidharan, Karthik. 2017. Field Experiments in Education in Developing Countries. In *Handbook of Economic Field Experiments*. Amsterdam: Elsevier pp. 323–385.

Muralidharan, Karthik, Abhijeet Singh and Alejandro Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review* 109(4):1426–1460.

Pritchett, Lant and Amanda Beatty. 2014. "Slow Down, You're Going too Fast: Matching Curricula to Student Skill Levels." *International Journal of Educational Development* 40:276–288.

Rockoff, Jonah. 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*, ed. Eric Hanushek, Stephen Machin and Ludger Woessmann. Amsterdam: Elsevier pp. 249–277.

Unesco, United Nations Educational, Scientific and Cultural Organization. 2019. "UNESCO Institute for Statistics Database." Online available, URL: `http://data.uis.unesco.org/` (last access: 04.12.2019).

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: Ranomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134(2):557–598.

A Appendix: Additional Analysis

A.1 Learning Gap and Grade Level Heterogeneity in our Sample

In order to examine the learning gap and grade level heterogeneity in our sample of primary school pupils, we convert their performance in the baseline assessment into a proficiency measure expressed in grade levels. As point of origin, we calculate for each participant her share of correct answers by item grade level. The score that a child obtains in our discrete proficiency measure is determined by those grade specific set of items, where the child scores at least 50% correct answers. In order to be assigned to a certain grade level, a participant needs to reach the 50%-threshold that corresponds with said grade level and all preceding grades. For example, a fourth grader that scored 80% on first grade items, 55% on second grade items and 40% on third grade items would be assigned to a second grade proficiency level. If a participant did not achieve 50% correct answers on first grade items, she is assigned to grade level <1.

Based on the previously specified measure, which is plotted in Figure 1b, we obtain a performance gap of two grades between the best and worst student in the *median* class of our sample. By construction the *mean* in the within-class performance range is lowest in third grade classes (about 1.3, i.e. the math abilities of students' within the same class cover on average 2.3 grades) and highest in sixth grade classes (about 2.4). A simple regression analysis also confirms that within-class variation is substantial, as classroom fixed effects only account for about 25% of the total variation at a certain grade level.

A.2 Attrition

In Table A.1 we examine whether the attrition at endline is correlated with the treatment status. To do so, we present results based on Linear Probability Models in columns (1) to (3), and on Logit Models in columns (4) to (6). The results unequivocal suggest, that the probability to miss the endline test did not depend on the treatment status.

Table A.1: Differences in attrition across treatments

Dependent var.: Attrition at endline	OLS			Logit		
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Lessons with Teacher	0.018		0.017	0.224		0.224
	(0.302)		(0.318)	(0.298)		(0.293)
T2: CAL-Lessons with Supervisor	0.021		0.026	0.263		0.330
	(0.203)		(0.115)	(0.202)		(0.116)
T3: CAL-Lessons with Teacher	0.023		0.025	0.280		0.315
	(0.226)		(0.190)	(0.215)		(0.173)
CX: Control Classes for Externalities	0.019		0.022	0.236		0.282
	(0.307)		(0.237)	(0.298)		(0.228)
Baseline math score		−0.002***	−0.002***		−0.024***	−0.024***
		(0.000)	(0.000)		(0.000)	(0.000)
Adjusted R^2	0.00	0.01	0.01	-	-	-
Pseudo R^2	-	-	-	0.00	0.02	0.02
Observations	3528	3528	3528	3528	3528	3528

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

A.3 Method of Inference and Robustness of our Results

As explained in Section 4.1, we apply two methods of inference. When we assess the impact of the different treatments relative to the children in pure control classes, the reported p-values are based on class-level clustered standard errors. Inference on within program school comparisons between the different treatments (including control classes subject to externalities), however, are based on a randomization inference test statistic with 2,000 random draws subject to the identical cut-off criterion as used in our re-randomization scheme.

This mixed estimation approach directly follows from our two-step randomization design (see Figure 2). Randomization inference is indispensable when comparing experimental groups within program schools since the underlying assignment process involved re-randomization. Conversely, selection of program schools and pure control schools was not based on re-randomization, making the use of randomization inference less critical.

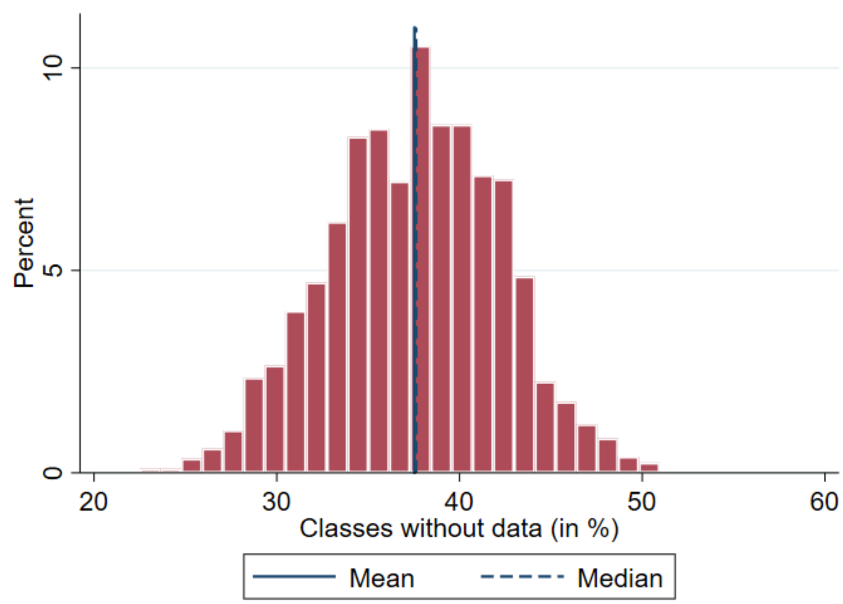


Figure A.1: Full re-randomization (incl. steps 1, 2a, and 2b) and the share of classes without data points (N=2000 draws).

Notes: This graph plots the distribution of the share of missing data points, when we conduct randomization inference by reiterating both stages of our randomization procedure. The large number of missing data points weakens the precision of our estimates, which explains why the p-values in the upper panel of Table A.2 increase by a factor of 5 to 10 compared to the p-values in Table 2.

While randomization inference is also preferable for assignment processes based on plain (or stratified) randomization (e.g. Young, 2019), its application is problematic in our case due to a particular feature of our study design: Out of the 162 eligible classes in pure control schools, we only collected data for a random sample of 40 classes. Implementing randomization inference for both stages of the randomization process thus comes with the downside that each draw will contain a considerable number of classes that did not participate in the assessments. As illustrated in Figure A.1, re-iterating the full randomization procedure yields an average of 37% of classes without data per draw. Even though missing data points in the replication procedure create an artificial loss of statistical power, we present the respective estimates as a conservative reference point.

Table A.2: ITT-Estimates on the effects of the different interventions on children's math scores with p-values based on clustered standard errors

	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teachers	2.904***	2.643**	0.165***	0.152**
	(0.005)	(0.012)	(0.006)	(0.013)
T2: CAL-Lessons with Supervisor	4.095***	3.869***	0.226***	0.214***
	(0.000)	(0.000)	(0.000)	(0.000)
T3: CAL-Lessons with Teacher	4.554***	4.328***	0.250***	0.238***
	(0.000)	(0.000)	(0.000)	(0.000)
CX: Control Classes for Externalities	2.595**	2.407**	0.147**	0.137**
	(0.011)	(0.017)	(0.013)	(0.020)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	1.191	1.226	0.061	0.063
p-value (β_{T8}=0)	(0.203)	(0.180)	(0.267)	(0.241)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	1.650*	1.686*	0.084	0.086*
p-value (β_{T9}=0)	(0.080)	(0.063)	(0.115)	(0.093)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.459	0.460	0.024	0.023
p-value (β_{T10}=0)	(0.606)	(0.599)	(0.637)	(0.636)
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: p-values based on traditional clustered standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

To assess the robustness of our results with respect to the method of inference, we report three versions of our benchmark analysis: In Table 2, the upper panel p-values are based on class-level clustered standard errors, while we run randomization tests in the lower panel. Table A.2 replicates these results, but inference is consistently based on class-level clustered standard errors. Finally, Table A.3 presents all results with p-values based on a full randomization tests.

Reassuringly, our main conclusion do not depend on the method of inference. When we apply traditional inference to the lower panel, as in Table A.3, changes in p-values are very small and do not show a clear pattern. And despite losing a lot of power when applying randomization inference to the upper panel, as in Table A.2, the only notable difference is, that the program externalities captured by β_{CX} turn insignificant with p-values around 0.13.

Table A.3: ITT-Estimates on the effects of the different interventions on children's math scores with p-values based on randomization inference

	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teacher	2.904*	2.643*	0.165*	0.152*
	(0.073)	(0.089)	(0.083)	(0.097)
T2: CAL-Lessons with Supervisor	4.095***	3.869**	0.226**	0.214**
	(0.009)	(0.013)	(0.015)	(0.018)
T3: CAL-Lessons with Teacher	4.554***	4.328***	0.250***	0.238**
	(0.006)	(0.006)	(0.007)	(0.011)
CX: Control Classes for Externalities	2.595	2.407	0.147	0.137
	(0.117)	(0.136)	(0.120)	(0.140)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	1.191	1.226	0.061	0.063
p-value (β_{T4}=0)	(0.214)	(0.194)	(0.268)	(0.244)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	1.650*	1.686*	0.084	0.086
p-value (β_{T5}=0)	(0.069)	(0.059)	(0.117)	(0.102)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.459	0.460	0.024	0.023
p-value (β_{T6}=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: p-values based on a two-sided randomization inference test statistic that the placebo coefficients are larger than the actual are shown in parentheses. The p-values were computed based on 2000 random draws.
* p<0.10, ** p<0.05, *** p<0.01.

B Appendix: Measuring and Converting Learning Outcomes

To measure math skills of third to sixth graders, we conducted two standardized math assessments during the school year 2018. Both assessments include 60 items and were designed as follows:

1. We summarized the Salvadoran math curriculum for grades 1–6 along the three topics (a.) number sense & arithmetic, (b.) geometry & measurement, and (c.) data & probability.

2. We then mapped test items from various sources on the Salvadoran curriculum. These sources are (a.) official text books of El Salvador, (b.) publicly available items from the STAR[17] evaluations in California, (c.) publicly available items from the VERA[18] evaluations in Germany, and (d.) exercises from the Swiss textbook MATHWELT.

3. We then gathered pilot data on 180 test items answered by 600 Salvadoran pupils in October 2017 and estimated the difficulty and discrimination parameters of test questions based on *Item Response Theory* (e.g. de Ayala, 2009).

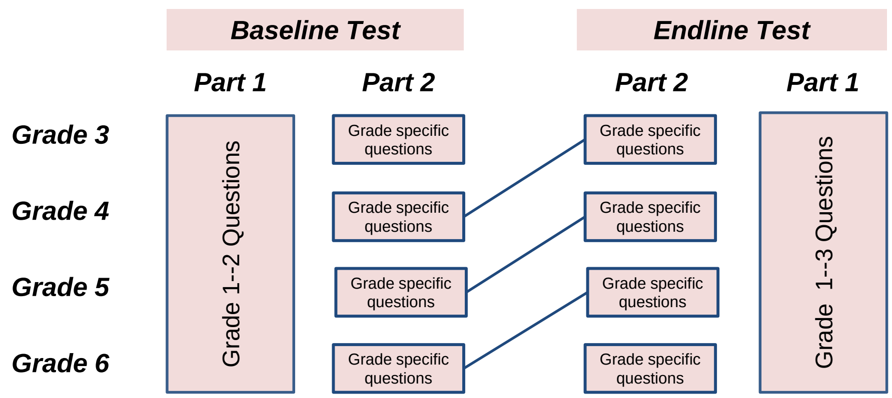


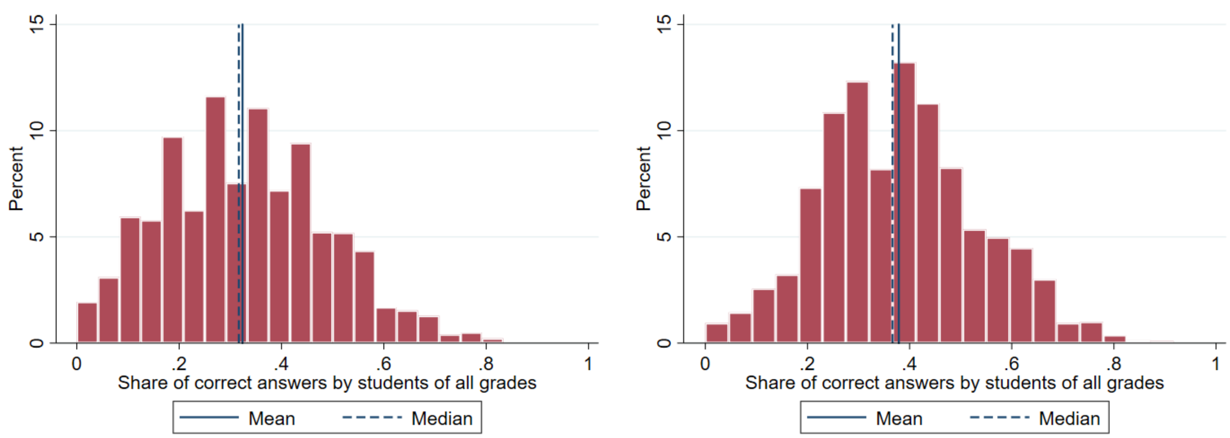
Figure B.1: Stylized illustration of the assessment design.
Note: Each part covers 30 items, adding up to 60 items per wave.

4. Finally, we designed paper and pencil maths tests using insights from step 3. The 60 items are selected such that they reflect the weighting in the official curriculum: 60–65% number sense & arithmetic, 30% geometry & measurement, 5–10% data & probability. Most items required a written answer, while the share of multiple choice questions varied between 10% and 15% depending on grade level. Figure B.1 illustrates how the math assessments at baseline and endline were structured and linked. Both assessments had two parts, with the first part being answered by all children independent of their grade. Moreover, the grade specific second part of 3rd/4th/5th graders in the endline assessment included many baseline questions of the 4th/5th/6th graders. This linking across grades and waves was essential to infer a commonly scaled ability score, i.e. the IRT scores.

[17]Further information on the Standardized Testing and Reporting (STAR) programme in California is available online: `www.cde.ca.gov/re/pr/star.asp` (last accessed: 14.01.2018).

[18]VERA is coordinated by the Institut für Qualitätsentwicklung im Bildungswesen (IQB), see `www.iqb.hu-berlin.de/vera` (last accessed: 14.01.2018).

Diagnostics. Table B.1 shows summary statistics on test items for each grade and wave of the assessment. In Table B.2 and Figure B.2, similar statistics are displayed for students' percentage scores. As can be seen, our test is not subject to relevant floor or ceiling effects: Hardly any students could not answer a single question on a given assessment and not a single student scored all items correctly. Similarly, only one item was not endorsed by anyone and no question could be answered by all students. On average, students gave correct answers to about 25-43% of the questions in a test booklet (column 2 in Tables B.1 and B.2). Figure B.3a shows the corresponding IRT-based test information function for the entire assessment, i.e. for all grades and waves combined (see below for details on IRT). As can be seen, our test is very informative for students across all ability levels. However, the assessment is skewed towards high difficulty levels, meaning that it allows to differentiate very precisely among high-achieving, but less precisely among low-achieving students. Ideally, the precision (or "information") of an assessment is highest around Theta = 0 where most students are located (see Figure B.3b). This implies that, on average, students should be able to answer about 50% of the test items. This reflects our decision to construct the assessment based on the official Salvadoran curriculum in spite of the mismatch between the curriculum and students' actual ability levels. Consequently, most of the included items could be answered by less than half of the students. While this curriculum-based approach allows for a more meaningful interpretation of results, it represents a slight loss in terms of test information. Nevertheless, sufficient questions of differing difficulty levels are covered to warrant the conclusion that, overall, our item battery provides a fairly reliable measurement instrument.



(a) Baseline test

(b) Endline test

Figure B.2: Distribution of percentage scores across students

Calculating IRT-Scores. Our math assessments allows us to project all outcomes on a common ability scale by using Item Response Theory. Instead of summing up the correct answers to a total score taken to represent a person's ability, Item Response Theory proposes a probabilistic estimation procedure. Ability is then viewed as a latent variable influencing the responses of each individual to each item through a probabilistic process: The higher a person's ability and the lower the difficulty of a particular test item, the higher the probability of a correct answer. In the simplest form of the model, the probability that individual i succeeds on item j can be expressed by the following function:

Table B.1: Item Diagnostic: The distribution of correct answers across items

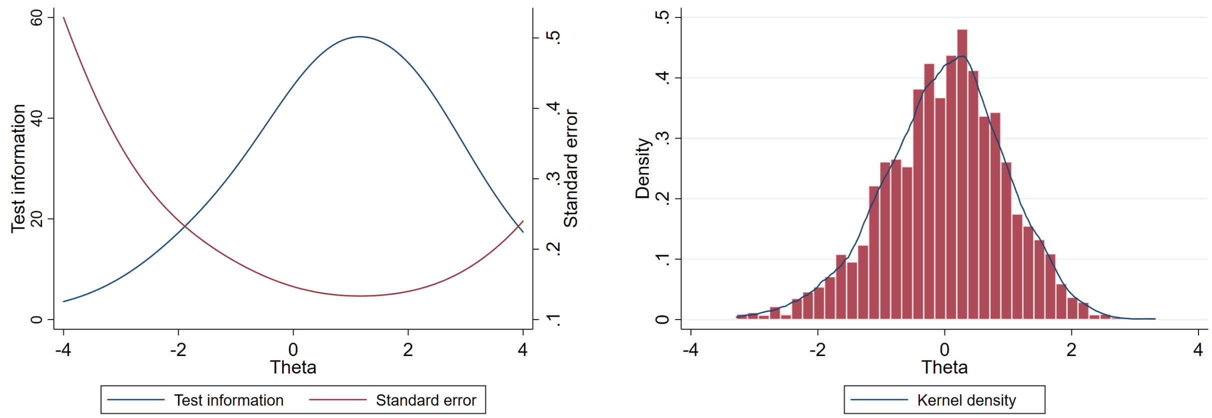
	Share of correct answers across items (in %)					
a. Baseline	Minimum	Mean	Median	Maximum	Share 0%[a]	Share 100%[b]
3rd Graders	0.4	24.9	18.3	87.3	0.0	0.0
4th Graders	2.4	30.9	25.5	94.2	0.0	0.0
5th Graders	0.4	34.9	26.6	96.6	0.0	0.0
6th Graders	0.4	38.7	27.4	96.4	0.0	0.0
b. Endline	Minimum	Mean	Median	Maximum	Share 0%[a]	Share 100%[b]
3rd Graders	0.9	34.1	23.5	95.8	0.0	0.0
4th Graders	0.5	36.0	31.0	98.0	0.0	0.0
5th Graders	0.0	38.9	32.3	98.8	1.7	0.0
6th Graders	1.3	42.6	37.2	98.9	0.0	0.0

Notes: The share of correct answers bases on those students that participated in both assessments, and hence constitute the main estimation sample. *a. Share 0%:* This column displays the share of items with zero correct answers. *b. Share 100%:* This column displays the share of items that were answered correctly by all test-takers.

Table B.2: Item Diagnostic: The distribution of percentage scores across students

	Percentage score across students (in %)					
a. Baseline	Minimum	Mean	Median	Maximum	Share 0%[a]	Share 100%[b]
3rd Graders	0.0	24.9	21.7	78.3	0.9	0.0
4th Graders	0.0	30.9	28.3	83.3	0.6	0.0
5th Graders	0.0	34.9	35.0	80.0	0.2	0.0
6th Graders	1.7	38.7	38.3	80.0	0.0	0.0
b. Endline	Minimum	Mean	Median	Maximum	Share 0%[a]	Share 100%[b]
3rd Graders	0.0	34.1	33.3	83.3	0.8	0.0
4th Graders	0.0	36.0	35.0	91.7	0.2	0.0
5th Graders	0.0	38.9	38.3	81.7	0.1	0.0
6th Graders	0.0	42.6	40.0	90.0	0.1	0.0

Notes: The distribution of percentage scores bases on those students that participated in both assessments, and hence constitute the main estimation sample. *a. Share 0%:* This column displays the share of students that answered zero questions correctly. *b. Share 100%:* This column displays the share of students that answered all questions correctly.



(a) IRT-based test information function

(b) Distribution of student abilities (Theta)

Figure B.3: Test information figure and distribution of students' abilities.

$$Pr(success_{ij}|b_j, \theta_i) = \frac{exp\{a(\theta_i - b_j)\}}{1 + exp\{a(\theta_i - b_j)\}}$$

with θ_i denoting the ability of student i, and b_j representing the difficulty of item j.

In this so-called *one-parameter model*, the probability that an individual endorses a particular item is thus a logistic function of the distance between the ability level of that individual and the difficulty of the item. Ability levels for each person and difficulties for all items can be computed through joint maximum likelihood estimation. IRT has many advantages over classical test theory. It tends to produce more reliable ability estimates, allows to link the scores of different individuals in different tests through overlapping items, and can help to better understand and improve the quality of a test (e.g. de Ayala, 2009).

As illustrated in Figure B.1 a selection of items overlap *(i)* between the baseline and endline assessments and *(ii)* across test booklets of different grades within an assessment wave. This allowed us to project the performance in the baseline and endline assessment onto a common scale through the estimation of an IRT one-parameter model. This procedure yields for every student i two ability estimates, namely one for the baseline assessment, i.e. θ_i^{BL}, and one for the endline assessment, i.e. θ_i^{EL}. The latter serves as outcome variable in the regression models that are labeled with "IRT-Scores".

Converting IRT scores to school year equivalents. To allow for an intuitive interpretation, IRT scores can be represented as school year equivalents. For this purpose, we re-scale ability estimates based on between-grade ability differences among pure control students at the time of the endline assessment; that is they are divided by the average difference between adjacent grades, which we calculated to be 0.31. That means, that the average ability difference between third and fourth graders, fourth and fifth graders, and fifth and sixth graders in October 2018 equaled 0.31.

The estimated program effects can then be interpreted as a proportion of the children's average progress during one school year. Note, however, that ability differences between grades 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 (e.g. literacy or science).