Expanding School Time and
the Value of Computer-Assisted Learning:
Evidence from a Randomized Controlled Trial 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

Evaluation Report[°]
on the CAL-Impact Program by Consciente
April 2019

Type of Program:	Additional math lessons, 180 min. per week
Version 1:	Computer-assisted learning lessons conducted by teachers i.e. CAL+TEACHER
Version 2:	Computer-assisted learning lessons conducted by supervisors i.e. CAL+SUPERVISOR
Version 3:	Math lessons instructed by teachers i.e. TEACHER
Target Country & Region:	El Salvador, Department of Morazán
Target Population:	Primary school children, grades 3 to 6
Evaluation Period:	April to October 2018
Evaluation Sample:	198 classes, about 3'500 children
Evaluation Design:	Randomized controlled trial (+ monitoring data & qualitative feedback)
Main Outcome:	Math ability measured via standardized assessments

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

Abstract

This report presents the results of a quantitative impact evaluation on a basic education program that aims to improve math skills among third to sixth graders in Morazán, El Salvador. The backbone of the evaluation is a randomized controlled trial (RCT) that evaluates the impact and compares the cost-effectiveness of three different interventions providing additional lessons in basic math, namely: *(i)* math lessons based on computer-assisted learning software and instructed by a qualified teacher, *(ii)* math lessons based on computer-assisted learning software and conducted by a supervisor, and *(iii)* math lessons taught by a qualified teacher but without learning software. We find that using computer-assisted learning software consistently outperforms the more traditional approach to teaching math. The computer-assisted learning courses not only produced a higher impact on math skills, but also turned out to be more cost-effective than traditionally designed math lessons. The evaluation also uncovers significant positive spillover effects in Consciente's partner schools: Children that were assigned to control classes in program schools experienced an increase in math knowledge compared to children in geographically separated control classes. Overall, this study confirms the benefits of expanding school time, especially if the additional math lessons are build around computer-assisted learning methods.

Contents

1 Introduction

Human capital is widely accepted to be a major driver of economic development. Investments in education are therefore a key ingredient to development strategies that aim at sustainably improving economic conditions in low- and middle-income countries. In recent years, considerable efforts by the international community and local governments fostered steeply increasing school enrollment rates in the developing world. In spite of this success, less than 20 percent of primary students in low-income countries pass minimum proficiency thresholds in reading and math, compared to 99 percent in high-income countries (World Bank, 2018, p.8). Children in developing countries evidently learn much less during their time spent in school than children in richer countries. In the light of these findings the World Bank (2018, p.3) concludes that "[s]chooling is not the same as learning". Without a doubt, understanding how educational programs should be designed and implemented is one key challenge for the international development community.

About the program Consciente strives to improve the quality of basic education in the rural department Morazán, El Salvador. With an average per capita income of 3.80 US-Dollars (USD) and an illiteracy rate of more than 20%, Morazán is one of the least developed regions in El Salvador (DIGESTYC, 2018). In this social context, the basic education program CAL-IMPACT aims at improving the math ability of primary school children of grades 3 to 6. As evaluated in this report, CAL-IMPACT provided additional math lessons for a period of six months (April – October 2018). These lessons were mainly taught in the afternoon to complement regular classes that are usually held in the morning.

Importantly, the program features three intervention arms with different inputs. Each of the three intervention arms comprises two additional lessons of 90 minutes per week, which almost doubled the beneficiaries' number of math lessons during the program phase. The first and second intervention arms are additional math lessons based on computer-assisted learning (CAL) software. Students work with the off-the-shelf platform *Khan Academy*, which allows them to learn at their own level and pace. The difference between the two interventions is that the additional computer lessons are either conducted by a temporarily contracted math teacher or by a temporarily contracted supervisor. *Supervisors* provide technical support but should not assist with math related questions. *Teachers*, in contrast, are also allowed to explain mathematical concepts, although students mainly work independently with the learning software. The third intervention arm offers more traditional math lessons instructed by contract teachers. These teachers are also hired by Consciente and teach refresher courses that repeat the math curricula of lower grades. Since the three interventions make use of different schooling inputs, their costs vary considerably. Abstracting from the market price of the computers, which were donated to Consciente, our ex-post calculation yields the following: With expenses of 53 USD per child, CAL lessons instructed by teachers are about 20% more expensive than both CAL classes conducted by supervisors (44 USD p. child) and more traditional math lessons instructed by teachers (45 USD p. child). From now on, we will refer to three intervention arms as *(i)* CAL+TEACHER, *(ii)* CAL+SUPERVISOR, and *(iii)* TEACHER.

Design of the impact evaluation   The evaluation of CAL-IMPACT primarily builds around a randomized controlled trial (RCT), which was specifically designed to measure changes in the beneficiaries' math ability that can be causally attributed to Consciente's work. RCTs are experimental evaluation designs that *randomly* assign eligible candidates to participate in the program or be part of the control group.[1] In our case this means that 118 school classes were assigned to one of the three intervention arms, while 80 classes were assigned to the control group. These control classes were not (directly) exposed to the NGO's work and therefore serve as a measure of how the beneficiaries' math ability would have developed absent the treatment. Random assignment is attractive because the beneficiaries and members of the control group share – absent the NGO's work – on average the same characteristics, for instance in terms of math ability. Metaphorically speaking, this allows us to compare apples with apples (instead of comparing apples with oranges) and hence warrants a causal interpretation of the obtained results. The RCT was designed to provide answers on the following questions:

1. How did the the three versions of CAL-IMPACT affect learning outcomes in basic math?

2. Which of the three program versions is the most cost-effective one? Or put differently: If the NGO aims at achieving maximum impact on learning outcomes per USD spent, which program version should it scale up?

The main data sources for the RCT evaluation are two *standardized math assessments*, conducted in February (i.e. baseline) and October (i.e. endline), as well as a *survey* on the participants' socio-demographic characteristics. This data is complemented with *monitoring data* collected throughout the school year as well *qualitative information* stemming from exchanges with the regional Ministry of Education, headmasters, regular teachers, participating children, and Consciente's local staff.

Main results   The evaluation shows that the additional CAL lessons had a bigger *impact* than the math lessons following a traditional approach: Our most conservative estimates suggest that an average participant of CAL+TEACHER gained math knowledge equivalent to about one-fifth school year, while the impact on those instructed by supervisors was about 20% weaker. The raw impact of additional math lessons without learning software was practically zero. Somewhat surprisingly, it appears that the presence of the NGO at schools had a sizable effect on learning outcomes, even among control classes that were not offered additional math lessons; we label these additional effects as "spillovers".

   In terms of *cost-effectiveness* both CAL interventions perform almost equally well, as the additional impact by teachers relative to supervisors is about proportional to the difference in implementation costs. Hence, if the NGO aims at achieving maximum impact on learning outcomes per USD spent, it could either choose CAL + TEACHER or CAL + SUPERVISOR; the former would yield a higher impact per child, while the latter allows for a larger scale given a certain budget.

---

[1]Critics of randomized controlled trials often point to ethical concerns related to random assignment and collecting data from control units. We address such concerns in Appendix C.

2 The local context

El Salvador is a lower middle-income country in Central America. At an international level, it is known for its high homicide rates. In 2015, the country gained sad notoriety as the "most violent place in the world" with 23 of the nation's 8 million inhabitants being murdered on an average day.[2] As to education, El Salvador's net primary enrollment rates are estimated at 85% and are thus slightly 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.[3]

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%, the district of Morazán ranks second-last among all Salvadorian districts in terms of educational attainment (DIGESTYC, 2018).
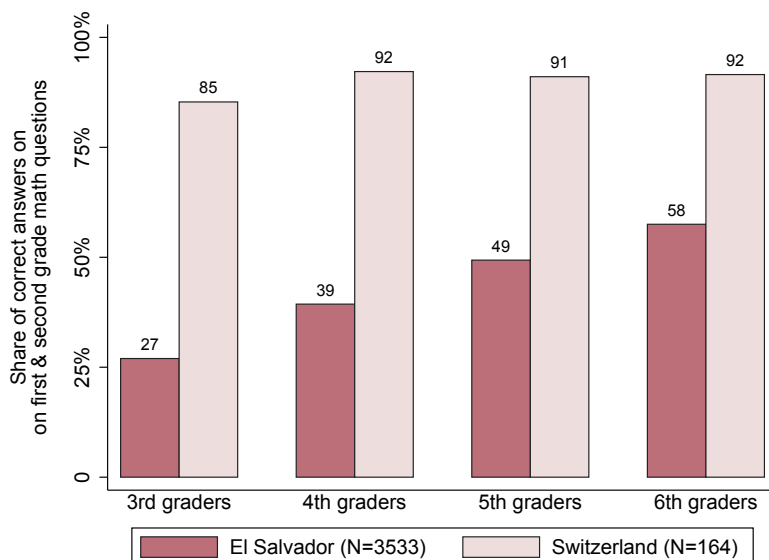
Figure 1: Average share of correct answers on first and second grade math questions among third to sixth graders in Morazán and in Switzerland.
Sources: Baseline data, February 2018.

Our math assessments with 3'533 third to sixth graders conducted in February 2018 further reveal, that primary school children barely grasp the most elementary concepts in math. Figure 1 shows that the share of correct answers to first and second grade questions increased from a dismal 27% among third graders to only 58% among sixth graders, who by then should have attended more than 1000 math lessons. To put these numbers into context, we conducted the same test with 164 children in Switzerland, who answered on average 85%–92% of the questions correctly.

---

[2]For detailed statistics on homicide see the online *homicide monitor* released by the Igarapé Institute, see `https://homicide.igarape.org.br/` (last access: 01.03.2019).

[3]Enrollment statistics according to the *World Development Indicators* provided online by the World Bank, see `https://data.worldbank.org/indicator` (last access: 01.03.2019)

Alarmingly, even the worst performing Swiss third grader outperformed the median sixth grader in Morazán.

These abysmal learning achievements raise questions about the underlying causes. While education in El Salvador faces a series of challenges, overstrained and poorly motivated teachers seem to be one major issue: Our data from school visits reveal high rates of teacher absenteeism so that on average 25% of lessons are canceled; this is a common issue in many developing country (e.g. Chaudhury et al., 2006). Moreover, our hiring assessments of (contract) teachers point towards a widespread lack of content knowledge. Many seemingly well-educated candidates, who entered our recruiting process, had trouble answering basic questions from the official grade 3 to grade 6 curriculum. For example, the pre-selected candidates – all of them certified teachers – had to compute the sum of two fractions (40% correct), or convert centimeters to meters (77% correct). On average, they answered only 50% of the grade 3 to 6 questions correctly. The worryingly poor content knowledge further mixes with outdated pedagogical techniques that basically follow the logic of "copy, learn by heart, and reproduce". Of course, those who cannot master the material themselves cannot pass it on. And those who have never experienced interactive teaching themselves cannot imagine that learning can also be exciting. This gives rise to a vicious cycle with poor education reproducing itself.

The Salvadoran Ministry of Education has recently shown considerable effort in addressing learning deficiencies in public schools. Until recently, primary schooling has been typically confined to morning lessons throughout El Salvador. The new SI-EITP (Sistema Integrado de Escuelas Inclusivas de Tiempo Pleno)[4] policy aims to extend school time over a full day and to complement traditional teaching with innovative learning approaches (MINED, 2013). The government not only hopes that longer schooldays boost learning outcomes, but that they further 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 innovative teaching and 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, mostly due to a lack of resources to pay for further teaching staff.

3   The CAL-Impact program

In this context, Consciente decided to launch a pilot for a computer-assisted learning project in primary schools. Adopting an evidence-based approach that is best summarized with the catchphrase "*innovate, test, then scale*", it was planned to thoroughly evaluate the pilot phase before (potentially) aiming at a continuous upscale.[5]

---

[4]Sistema Integrado de Escuelas Inclusivas de Tiempo Pleno=Integrated System of Inclusive Full Time Schools.

[5]*Innovate, test then scale* seems like an obvious approach, but is in fact a radical departure from what has been the standard practice in development cooperation. International pioneers pushing towards an evidence-based approach haven been *J-PAL*, a US based research center founded in 2003, and the *International Initiative for Impact Evaluation (3ie)* launched in 2008. Proponents of an evidence-based approach in the Swiss development cooperation include Günther (2016) and Kudrzycki and Günther (2018).

Why CAL? The choice in favor of a computer-assisted learning approach followed from *(i)* an analysis of the deficiencies in Morazán's educational services described in Section 2, *(ii)* the approach's good alignment with national policies and *(iii)* insights from rigorous impact evaluations of educational programs in developing countries:

(i) Software-based lessons offer two decisive advantages: *First*, they provide sophisticated learning videos, developed by educational experts and following proven pedagogical concepts. In the light of the teachers' insufficient content knowledge, the software serves as a substitute for qualified teaching staff. *Second*, if PC labs are furnished with one PC per child, pupils can cover basic math concepts at their own pace, which may prove valuable, especially in large and/or heterogeneous classes.

(ii) As mentioned above, the Salvadorian government has recently pushed for a strategy that aims to expand school days to the afternoon and promote learning through innovative teaching. One key design parameter was to align the program with this national policy to secure the support and collaboration of the Ministry of Education. Indeed, both core elements of the program, i.e. afternoon lessons and equipping schools for computer-assisted learning courses, were well received and facilitated the co-coordination with local authorities.

(iii) The computer-assisted learning approach not only offers a convincing narrative for an education intervention in El Salvador, it has also proven highly effective in several impact evaluations, especially for teaching math in developing countries.[6] Figure 2 summarizes the findings from rigorous evaluations that assessed the impact of CAL-lessons and provisions of computers on the math learning outcomes of primary school children. It shows that simply equipping primary schools with computers does not raise learning outcomes. When the computers are specifically used for computer-assisted learning in math, however, they have demonstrably improved learning outcomes in several evaluations and contexts. Nonetheless, a number of open questions remain, which are relevant both from an applied and research perspective: On the one hand, these evaluations only used control groups that did not receive additional math lessons and hence identify the impact of *additional* math lessons based on computer assisted learning, rather than the value added through the CAL component. On the other hand, all these evaluations examined interventions that used customized software specifically designed for the local context. These are either expensive or unavailable to NGOs working in other language regions. Hence, a second question is, whether off-the-shelf software, like *Khan Academy* (available in 19 languages), can also be effectively used in the classroom.

---

[6]For a systematic review of the accumulated evidence on the effectiveness of different educational interventions we refer to the recent and comprehensive report by the International Initiative for Impact evaluation (3ie), see Snilstveit et al. (2015).
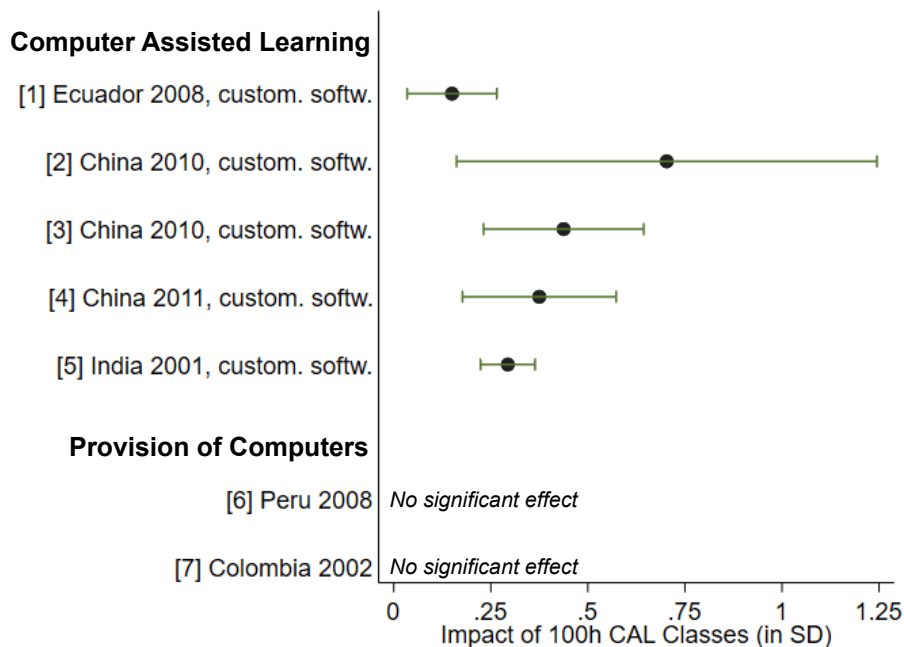
Figure 2: International evidence on the impact of technology-assisted interventions.
Sources: [1] Carrillo, Onofa and Ponce (2011), [2] Lai et al. (2015), [3] Yang et al. (2013), [4] Mo et al. (2014) [5] Banerjee et al. (2007), [6] Cristia et al. (2012), [7] Barrera-Osorio and Leigh (2009)

Main features of CAL-Impact After settling on a CAL-based approach to improve math abilities among primary school children, Consciente teamed up with economists and sociologists from the University of Bern to work out the specifics and simultaneously design the experimental impact evaluation. The early collaboration with the evaluation team proved very valuable, because it allowed to purposefully attune CAL-IMPACT and its evaluation.

The design phase surfaced many unanswered questions, which involve cost-relevant parameters on the side of the implementer, for instance: How many lessons per week should be offered, or in an economist's jargon: at what exposure do decreasing marginal returns kick in? Should children share a computer or is it worthwhile to provide one workstation per child? What complementary inputs would have the potential to boost the program's cost-effectiveness?

Finally, CAL-IMPACT was designed such that Consciente could gain substantiated insights on the following questions:

1. Is it worthwhile to ship IT-equipment to El Salvador and build up capacity for its long-term maintenance? Or would it be more cost-effective to focus on interactive math lessons that repeat the most basic concepts without computers?

2. Should CAL-lessons be instructed by certified primary school teachers or would it be sufficient to hire less qualified staff as supervisors who primarily provide technical assistance?

To answer these questions, CAL-IMPACT features three intervention arms that are illustrated

in Figure 3. Each of the three intervention arms comprises two additional lessons of 90 minutes per week, which almost doubled the beneficiaries' number of math lessons during the program phase. The first and second intervention arms are additional math lessons based on computer-assisted learning software, while the third intervention arm comprises additional math lessons instructed by a teacher without using software. Comparing the impacts and costs across the three intervention arms allows to gain insights on the practical value of using computers and the complementarity/substitutability between teaching skills and learning-software.

Version 1: *CAL + Teacher*	**Version 2:** *CAL + Supervisor*	**Version 3:** *Teacher*
2 x 90min p. week	2 x 90min p. week	2 x 90min p. week
39 classes, ~800 children	39 classes, ~800 children	40 classes, ~800 children
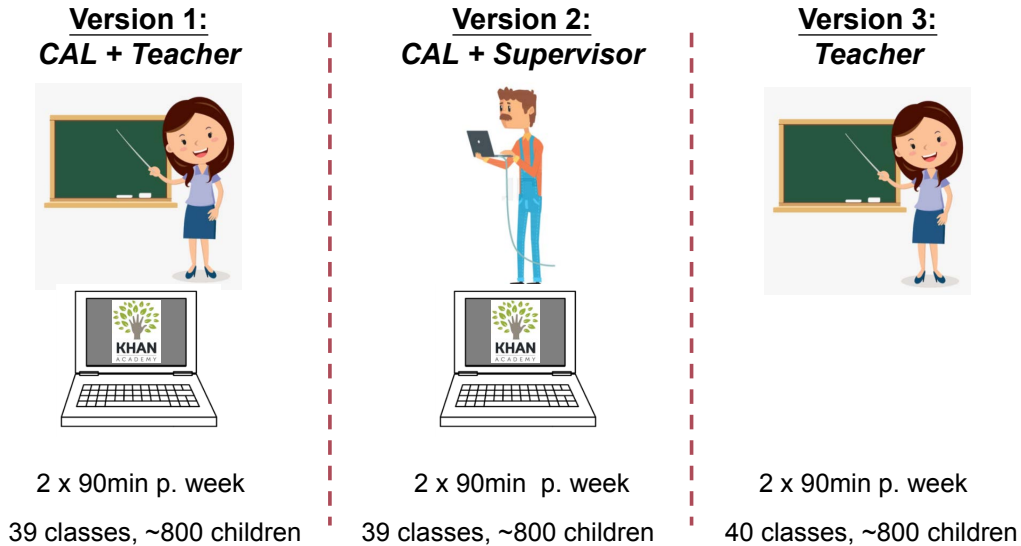
Figure 3: An illustration of the three program version of CAL-IMPACT

The *CAL-lessons* in the first two intervention arms 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 high-quality 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 instructions 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 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 third intervention arm that features more traditional *math lessons instructed by a teacher*. According to their initial ability level, children were arranged in two different table groups where they worked on level-appropriate plans. 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*, in all treatments individualized learning was combined with educational games. For this purpose, a manual containing animation, concentration and math games was developed. It shows simple techniques to promote students' collective learning as well as their motivation. Games were usually played at the beginning or at the end of each session. While supervisors made only use of animation and concentration games, teachers were additionally instructed to play 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 mathematical assessment and a job interview.[7] They were employed on short-term contracts and paid 300 USD per month for assuming four classes.[8] For lessons that were not conducted, teachers received no remuneration. To make effects between the treatments more reliably comparable, all teachers were assigned an equal number of CAL and non-CAL lessons. Before the intervention, teachers were intensely trained in the use of the software and the application of the educational games, and central pedagogical strategies and mathematical concepts were reviewed. Teaching was tightly monitored by Consciente through monthly meetings and unannounced classroom visits during the duration of the project.

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 not required to have any teaching credentials or a particular educational degree and were selected based on a job interview. As a prerequisite they only needed to have minimal IT skills and some experience in dealing with children. 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.

4 Evaluation design and randomization scheme

Evaluation design   The evaluation of CAL-IMPACT builds around a RCT to identify the causal impact of the three different interventions. Figure 4 illustrates the logic behind RCTs and why they allow us to directly measure the program's impact: Participants randomly assigned to the program share on average the same characteristics as the candidates randomly assigned to the control group. For instance, we would expect that they score about equally well in a math assessment conducted before the program is implemented; that is why the average math score of the treatment (i.e. rose

---

[7]As pointed out in Section 2, the content knowledge of applicants was generally poor. Since a sizable proportion of the applicants had to be selected, mathematical knowledge of those who got the job was unsatisfactorily low. An assessment with about 231 regular grade 3-6 math teachers in Morazán suggests that project teachers have slightly better performance levels as normal math teachers.

[8]This corresponds to 8x90 minutes of teaching per week, or – including preparatory work – to a 60% job. A smaller group of teachers only assumed two classes (i.e 4x90 minutes of teaching per week, or approx. a 30% job).
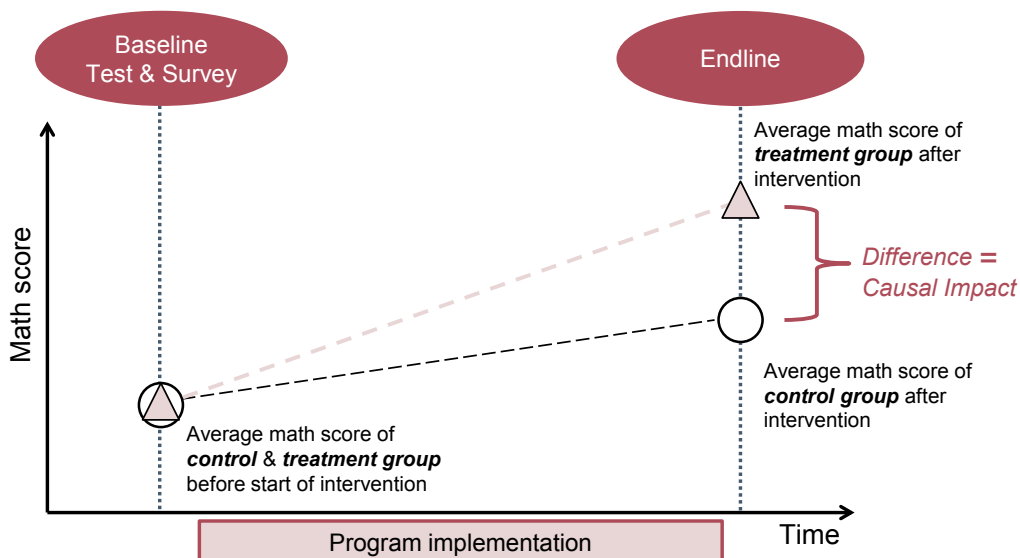
Figure 4: Stylized illustration of how a RCT isolates the causal impact of the evaluated program.

triangle) and control group (i.e. hollow circle) at baseline are drawn on the same spot in Figure 4. As two rounds of data are collected, the first before the intervention starts (i.e. the baseline survey) and the second after it finishes (i.e. endline survey), one can actually test this assumption. If it holds, the endline data then allows us to directly infer the causal impact of the program: Since the control and treatment groups share – absent the program – the same characteristics, any differences observed after the implementation of the program can be attributed to the work of the NGO.

This evaluation exactly followed this logic: As the stylized timeline in Figure 5 illustrates, the evaluation started in February 2018 with a baseline assessment and a survey covering all control classes and program classes. The intervention offering additional math classes started in mid-April 2018 and was 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.

Randomization scheme   Random assignment has been mentioned several times now. As the mechanism implemented in this evaluation has several layers, see Figure 6, we carefully walk through them step by step.

Starting point are the 306 primary schools in Morazán. Limited financial resources did not allow Consciente to implement Cal-Impact in all these schools. Due to this factual oversubscription, the evaluation can make use of a control group without lowering the number of beneficiaries reached by the NGO (also see Appendix C). In a first step, Consciente together with the evaluators defined eligibility criteria for a *preselection* of primary schools. These criteria included the following:

- **School size**, *eliminates 225 schools:* A school was considered too small, if it had integrated

---

[9]The school year in El Salvador starts in mid-January and ends in November.
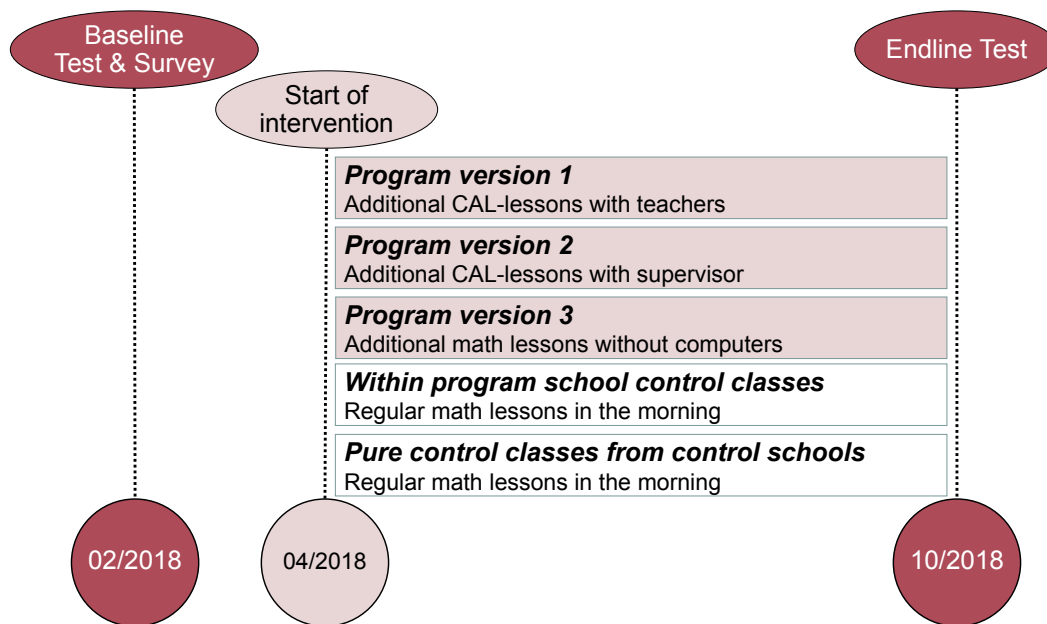
Figure 5: Timeline of the CAL-IMPACT evaluation.

classes (across grades) or gaps in their 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.

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

- **Accessibility**, *eliminates 7 of the remaining 67:* Extremely remote schools that are hardly accessible by car were discarded; this was based on an assessment of the local NGO-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 (see Figure B.1a in the Appendix). However, project resources still did not allow Consciente to operate in all 57 schools. Therefore, in *randomization stage 1*, 29 of the 57 schools were randomly chosen to be part of CAL-IMPACT (see Figure B.1b in the Appendix). To increase the precision of our estimates, the assignment was stratified by school size, local population density and students' access to a computer room; stratification means that schools were grouped by characteristics before randomly selecting one half of each group into the CAL-IMPACT program.

10

Figure 6: Sampling and randomization scheme.

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 ar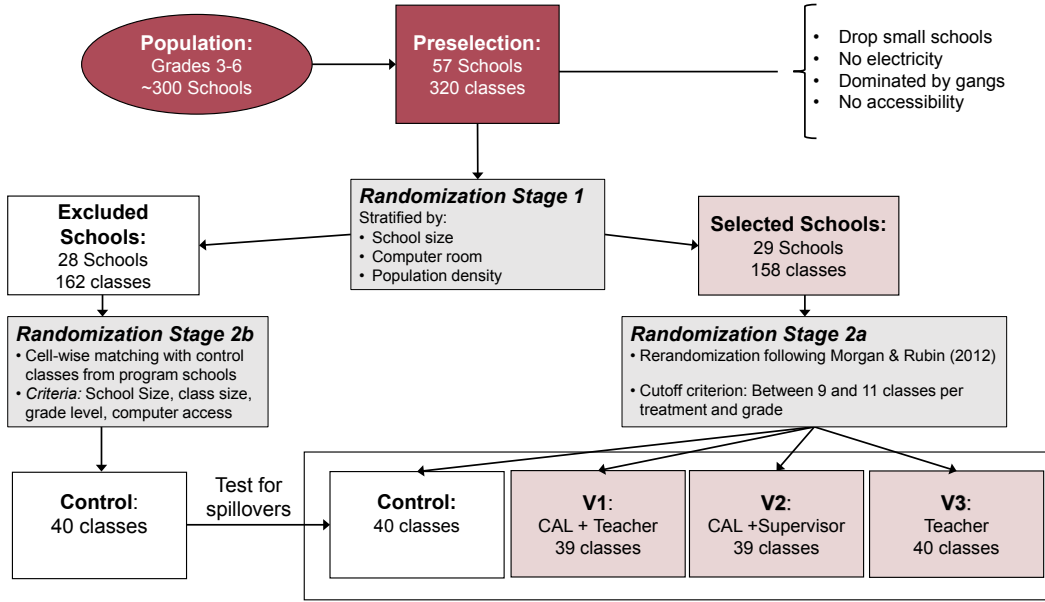ms (see Figure B.2a in the Appendix). Following Morgan and Rubin (2012) we re-run the randomization routine until the interventions were balanced across schools and grades. This mechanism assigned 39 classes to CAL+TEACHER, 39 classes to CAL+SUPERVISOR, 40 classes to TEACHER, and 40 classes to the control group.

In terms of cost-efficiency, class-level randomization has a decisive advantage over school-level randomization. However, this upside comes at the cost of greater risk of spillovers effects. Spillover effects occur if an intervention does not only affect individuals in the treatment group but also those in the control group, which would yield biased effect estimates. Since students within schools are more likely to influence each other than students from different schools, interventions which might produce substantial treatment externalities usually rely on school level randomization (e.g. Miguel and Kremer, 2004). Given the nature of the intervention, treatment externalities were initially expected to be small. Nonetheless, in *randomization stage 2b*, 40 additional control classes from non-treatment schools were included in the study to estimate the size of potential spillover effects (see Figure B.2b in the Appendix). As these additional "pure" control classes are geographically separated from the intervention, an thus unlikely to be affected by Consciente's work.. The *pure control classes* were randomly selected from the 28 control schools by cell-wise matching them 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 that are systematically compared in the evaluation, namely the 118 classes assigned to the three different intervention groups, 40 control classes from the 29 program schools, and 40 pure control classes from the 28 control schools.

5 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) monitoring data was recorded during unannounced school visits throughout the program phase, and (iv) qualitative information was collected in interviews with different stakeholders and via an online-questionnaire.

The *math assessments* cover the primary school curriculum of El Salvador. The weighting of questions across the three main topics (a) number sense & elementary arithmetic ($\sim$65%), (b) geometry & measurement ($\sim$30%), and (c) data & statistics ($\sim$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 were inspired by El Salvador's official textbooks and various international sources of student assessments. Hence, the skills measured with these tests cover broad aspects of basic math; the appendix section B.3 explains their design step by step.

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 labeled as Item Response Theory (e.g. de Ayala, 2009). This means that we can directly compare children across grades and express their learning gains between baseline and endline assessment in terms of how many additional school years would be required to reproduce the same effect. The conversion of our main regression estimates into program effects measured in terms of additional school years is explained in the Appendix B.3.

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 the children to answer them.

From May to September, NGO staff made on average five unannounced school visits to collect *monitoring data*. They covered both regular lessons as well as CAL-IMPACT lessons and collected data on teacher attendance, student attendance, computer usage, and the implementation of the additional math lessons in the afternoon. These unannounced visits established a feedback loop to the local program management, and allow the evaluators to examine potential channels causing spillover effects between program classes and control classes.

Finally, we further gathered *qualitative feedback* from the different stakeholders involved in the project. While all teachers and school directors participated in evaluation meetings and an extensive online survey, information on the perceptions of other actors was collected through (unstructured) conversations and interviews.

Balance at baseline   As explained in connection with Figure 4, random assignment to the three intervention arms and the two control groups is designed to result in five groups of primary school

classes that share very similar characteristics prior to the intervention. Table B.1 in the Appendix shows summary statistics for the main variables collected before the program started, and formally tests the hypothesis that the groups are – statistically speaking – alike: The table displays means and standard errors for three different math scores, sociodemographics of students, class room variables and school variables separately calculated for the three program and two control groups. Most importantly, the comparison of characteristics between the five groups shows differences that remain aloof from the 0.1-threshold for statistical significance. We can therefore conclude that randomization was successful, and that we will indeed compare apples with apples when calculating the impact of three program arms. Moreover, the results from the statistical models discussed in the next section absorb all variation in outcome math scores that can be attributed to (negligibly) varying group characteristics rather then the implementation of Cal-Impact.

6   Results

6.1   The effect of CAL-Impact on learning outcomes

How did the the three versions of Cal-Impact affect learning outcomes in basic math? This section answers this question based on an in-depth analysis of the math assessments conducted at the end of the pilot-phase in October 2018. Since we learned from Table B.1 that the two control and three program groups shared on average the same characteristics at baseline, any differences we can observe in the endline data can be attributed to the NGOs work. This section presents the main results on the basis of graphical illustrations that summarize the technically more detailed analyses discussed in the Appendix A. All the impact estimates are plotted on a scale showing school year equivalents; Appendix B.3 explains how we convert our quantitative estimates to obtain that scale.

Figure 7 plots the impact estimates for the three program versions. These estimates include all students that were originally assigned to one of the three intervention arms, irrespective whether they actually attended the additional classes or not.[10] Panel (a) uses control classes from within program schools, while Panel (b) is based on a comparison with pure control classes from control schools.

What are the main takeaways from these two graphs? *First*, they reveal a distinct ranking in terms of impact across the three interventions: With a gain in math ability equivalent to 0.19 – 0.47 school years, CAL-lessons instructed by teachers had the strongest effect on learning outcomes. CAL-lessons conducted by supervisors performed slightly weaker (0.14–0.43), while additional math lessons without CAL software clearly showed the smallest impact ranging between 0.05 and 0.31 additional school years. *Second*, when comparing classes within program schools, only the effect of CAL-classes instructed by teachers is statistically significant. The average impact of CAL-classes conducted by supervisors just falls short of crossing the 10%-threshold, although this depends somewhat on the specification used (see Appendix, Table A.1). The impact of additional math classes is statistically negligible, at least when doing a within school comparison. *Third*, we briefly

---

[10]The technical term used in the social sciences for this type of estimates is *Intent to Treat (ITT)* estimates.

(a) **Lower bound excl. spillovers**
Program school control units, 90% confidence intervals.

(b) **Upper bound incl. spillovers**
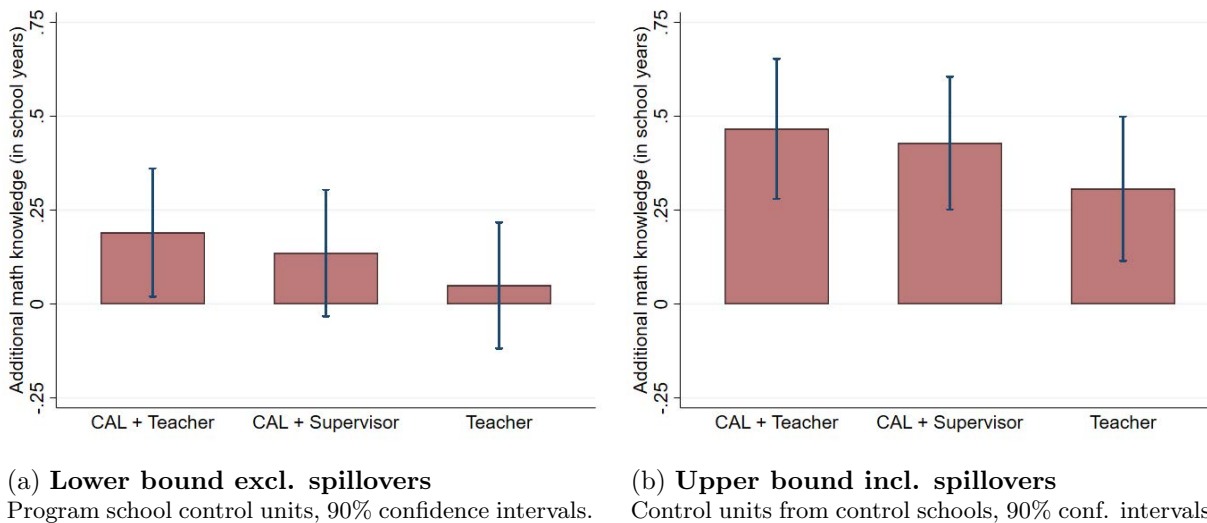Control units from control schools, 90% conf. intervals.

Figure 7: Average impact of the three intervention arms on math skills measured in school year equivalents. Impact estimates in Panel (a) are net of spillovers, while Panel (b) includes them.
Source Panel (a): Column (6) in Table A.1 normalized by factor 0.46, see Appendix B.3.
Source Panel (b): Column (6) in Table A.2 normalized by factor 0.46, see Appendix B.3.

discussed the possibility of spillover effects from treatment classes to control classes *within* the CAL-IMPACT program schools. Indeed, the differences in impact estimates when using control classes from program schools (i.e. Panel a) and external control classes from control schools (i.e. Panel b) are substantial. While Panel (a) displays lower bound estimates that net out any spillovers effects, Panel (b) shows upper bound impact estimates that fully attribute potential spillovers to the three CAL-IMPACT program versions.

Spillovers From a scientific point of view spillovers usually are an issue, because they confound the causal interpretation of the estimated impact effects. Since we were aware that spillovers within program school might occur, forming a geographically separated control group was important to validate the results. From a policy perspective, the positive spillovers that we observe when comparing Panel (a) and Panel (b) of Figure 7 are of course attractive as they increase the aggregate effect of the NGO's work.

In the light of these results, one is inclined to ask, how the work of Consciente in program schools actually exerts a positive impact on those classes not directly exposed to the three interventions. Many potential channels could be at work, and our (monitoring) data does not allow us to conclusively examine these channels.

For instance, the presence of the NGO and the installation of computer rooms for the additional afternoon lessons may have increased the *motivation of regular staff and children.* Indeed, most schools were highly pleased about their selection for project participation (and particularly about the reception of the IT equipment) and eager to make a good impression, hoping they would continue to be part of the intervention after the end of the pilot phase. The daily presence of NGO staff may not only have boosted teacher motivation, but could also have lead to increased performance

14

pressure. While we cannot directly examine this channel, one indication backing this hypothesis would be lower cancellation rates in regular lessons of program schools compared to cancellation rates in control schools. However, the analysis of cancellation rates based on our monitoring data (see Table A.5 in the Appendix) does not reveal significant differences supporting this claim. Similarly, we do not find any evidence that the attendance rate of children in regular lessons varies with their school's or class's assignment to the program (see Table A.4 in the Appendix).

Beside such motivational considerations, several other channels come to mind: Despite countermeasures, children of control classes in program schools might have used Consciente equipment; or they became inspired to use the freely available learning software in their freetime. Moreover, children from control classes might have benefited through social learning from their peers and/or siblings that participated in CAL-IMPACT classes. Finally, competition effects, sometimes labeled as "John Henry Effect", could partly explain the differences, for instance if children from control classes in program schools were more motivated to outperform their peers on the test day than children in control schools, where typically only one class had to take the assessment. While our evaluation cannot provide a definite answer on the source of the spillovers, our approach to work with two control groups, one within program schools and one in control schools, is well suited to develop further insights on this matter.

6.2 Student attendance in CAL-Impact lessons

One of the most important monitoring indicators that has been collected throughout the program phase is student attendance in the CAL-IMPACT lessons. Figure 8a plots this data: the ten bins cover the categories *0–10% attendance* up to *90–100% attendance* and their height reflects the percentage share of students belonging to each category. It also displays the mean attendance (solid vertical line) and median attendance (dashed vertical line) across eligible students.
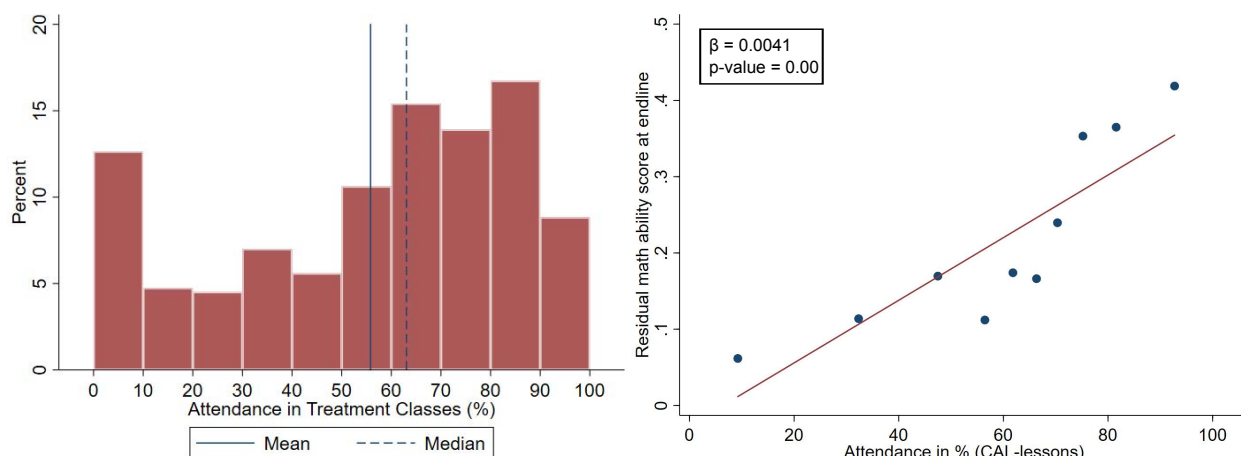
(a) Attendance in CAL-IMPACT math classes   (b) Correlation btw. attendance & endline math score.

Figure 8: Attendance in additional math classes and its correlation with the endline math score.
Note: Panel (b) shows the partial correlation net of the baseline math score, student characteristics, and school fixed effects. Each dot represents the average attendance rate and average residual math score of about 250 children.

Attendance by students was rather a weak spot of CAL-Impact. Monitoring data shows that only 11% of CAL-IMPACT classes were canceled, which is substantially below the 25% cancellation rate of regular classes held in the morning. At the same time, the average attendance rate of students in the additional math lessons, which is about 60%, compares unfavorably to the 90% attendance rate in the regular lessons. About 15% of eligible children even attended less than 10% of the additional math lessons, and therefore could hardly benefit from the program.

Judging from feedback by teachers and headmasters, one handicap is that the additional lessons are scheduled in the usually school-free afternoon. Although Consciente offered free lunch to the children attending afternoon classes, three issues remain: *First*, some students live in remote places that are not serviced by public transport in the afternoon. *Second*, some parents were concerned that their children return home too late, especially if their route to school crosses gang-dominated neighborhoods. *Third*, some parents mentioned that they prefer their children to help at home rather than to spend the afternoon in school.

(a) **Lower bound excl. spillovers**
Program school control units, 90% confidence intervals.

(b) **Upper bound incl. spillovers**
Control units from control schools, 90% conf. intervals.
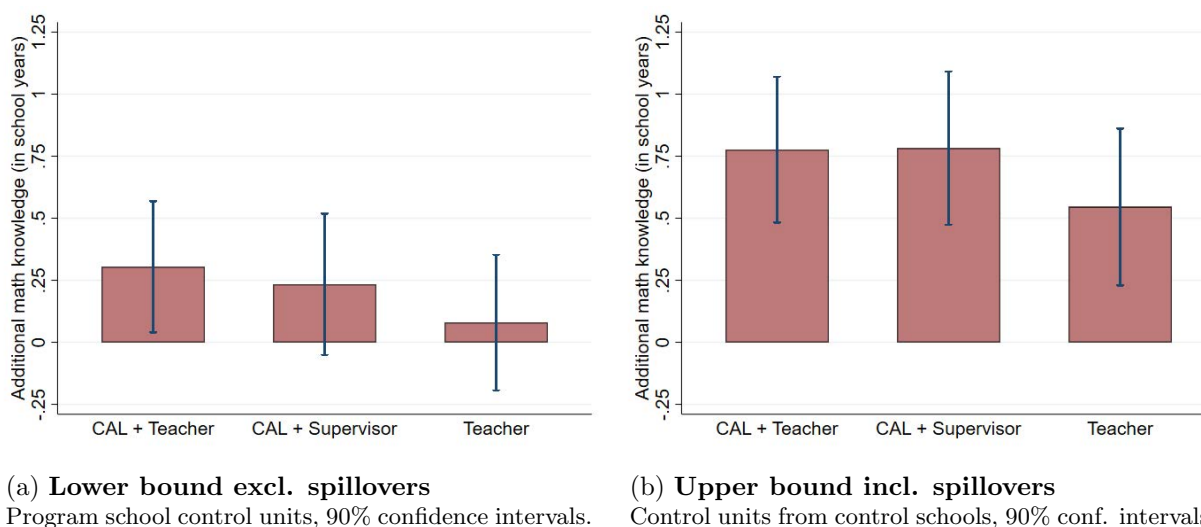
Figure 9: *Hypothetical scenario with full attendance:* Impact on math skills measured in school year equivalents. Estimates in Panel (a) are net of spillovers, while Panel (b) includes them.
Source Panel (a): Column (6) in Table A.6 normalized by factor 0.46, see Appendix B.3.
Source Panel (b): Column (6) in Table A.7 normalized by factor 0.46, see Appendix B.3.

Should this be a concern to the NGO? The data suggests that the modest attendance rates considerably mitigate the program's overall impact. Figure 8b plots the correlation between attendance in the two CAL-math classes and the math score at endline when controlling for any observable student characteristics including the baseline score. As one would expect, the strongly positive correlation shows that those attending the additional math classes experienced more learning gains than those missing most classes.

By how much would the program's overall impact increase if Consciente could boost the attendance rate? Assuming that the effects of the program are the same across students, we can estimate the average impact of the program under a *hypothetical full attendance scenario*.[11] Figure 9 plots

---

[11]In technical terms, we estimate an *instrumental variable (IV) model* which is discussed in section A.3.

these hypothetical program impact estimates if the eligible students had consequently attended all additional math lessons. A comparison with the the actual program impact plotted in Figure 7 suggests that the effect estimates increase by about 50%. While the strong assumptions underlying this exercises call for a cautious interpretation, the results suggest that the overall impact of the program could be substantially boosted if the NGO finds ways to increase the attendance rate among the beneficiaries. With full attendance and when including spillovers, the estimated learning gains from additional CAL-lessons increase to the equivalent of 0.75 school years.

6.3   A cost-effectiveness comparison

So far, the discussion on the impact of the different intervention arms completely ignored the cost side. From a policy perspective, however, learning about the relative cost-effectiveness is probably more relevant than comparing raw impact estimates across programs. We therefore turn to the second main question of the evaluation: If Consciente aims at achieving maximum impact on learning outcomes per USD spent, which program version should it scale up?

To answer this question, one needs accounting data to calculate the costs per child for each intervention arm. In the case of CAL-IMPACT, two items carry most of the weight: First, salaries of teachers (300 USD per month) and supervisors (180 USD per month) represent about 45% of the program costs. Second, furnishing schools with the essential IT-equipment and providing technical maintenance account for about 20% of the expenses. It needs to be pointed out, however, that the lion's share of the IT-equipment was donated to Consciente, substantially lowering the relative costs of the CAL-based interventions compared to a scenario without such in-kind contributions. Since, according to Consciente, in-kind donations can be acquired more easily than funds, we use the factual program expenses for our cost-effectiveness calculations instead of putting a price tag on the donated equipment. Some equipment, for instance monitors, had to be bought, and several one-time costs incurred related to shipping and customs duty. We add these one-time costs and the actual purchase costs together in order to infer the average fixed costs per workstation. As we assume the equipment's life span to be 4 years, annual depreciation rates of 25% are used when calculating yearly program costs. This yields the following unit costs per school year:

- CAL+TEACHER: 53 USD per beneficiary

- CAL+SUPERVISOR: 44 USD per beneficiary

- TEACHER: 45 USD per beneficiary

Having estimated the impact of each intervention arm and knowing their unit costs, one can make the following thought experiment: Assume that the NGO has additional funds available and wants to scale up one of the interventions. It can either choose the expensive but also most effective CAL + TEACHER intervention or select the cheaper but also less impactful CAL + SUPERVISOR or TEACHER option. Hence, the NGO faces a trade-off between having a larger impact per child or targeting more children. But what can we learn from combining the per unit program costs with

17

(a) **Lower bound excl. spillovers**
Program school control units, 90% confidence intervals.

(b) **Upper bound incl. spillovers**
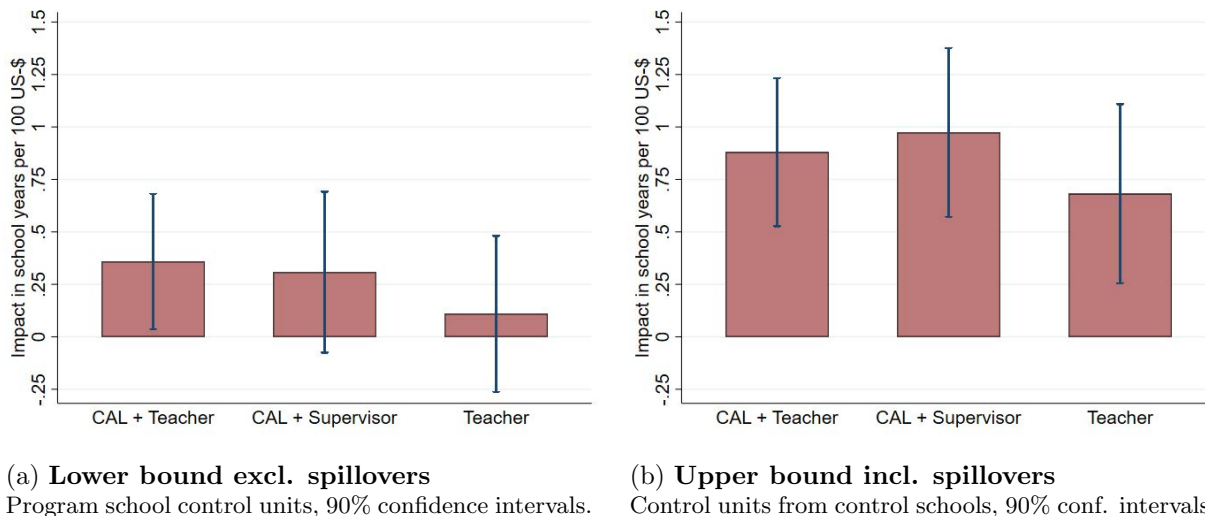Control units from control schools, 90% conf. intervals.

Figure 10: Average impact of the three intervention arms on math skills in school year equivalents. Impact estimates in Panel (a) are net of spillovers, while Panel (b) includes them.
Source Panel (a): Column (6) in Table A.1 normalized by the factor 0.46 and the annual unit costs.
Source Panel (b): Column (6) in Table A.2 normalized by the factor 0.46 and the annual unit costs.

the impact estimates? One can calculate the learning gains the NGO can "buy" with every 100 dollars it spends on a particular program. This measure allows the NGO to make an evidence-based decision about what program option would maximize the total impact of its future investments.

We normalize the impact estimates from Figure 7 with the unit costs presented above, and then plot the potential learning gains per 100 USD invested in Figure 10. We gain two key insights from this exercise: *First*, shipping the IT-hardware to El Salvador and setting up computer labs payed off: Both CAL-based intervention achieve a higher cost-effectiveness than the more traditional math lessons instructed by certified teachers. *Second*, the two CAL interventions are similarly cost-effective: When excluding spillover effects in Figure 10a, CAL-lessons conducted by teachers marginally outperform CAL-lessons instructed by supervisors, but this ranking reverses when including spillover effects, as plotted in Figure 10b. Apparently, the additional impact of teachers instructing CAL-lessons is offset by the proportional increase in the implementation costs due to their higher compensation compared to supervisors. The estimates quantify the learning gains for every 100 USD invested at an equivalent of about 0.3 (excl. spillovers) to 0.8 (incl. spillovers) school years.

Putting the results into context One of the most striking patterns in the data is the relatively poor performance of additional math lessons compared to the computer-assisted courses. During the recruiting process of contract teachers, the applicants participated in a short assessment on basic math concepts. As we previously discussed in Section 2, the fragmentary content knowledge of the seemingly well-qualified candidates was a surprise to the NGO and a source for concern throughout the project's implementation. Field visits in schools strengthened the impression that some teachers, both among the regular faculty and Conciente's staff, struggle with the concepts
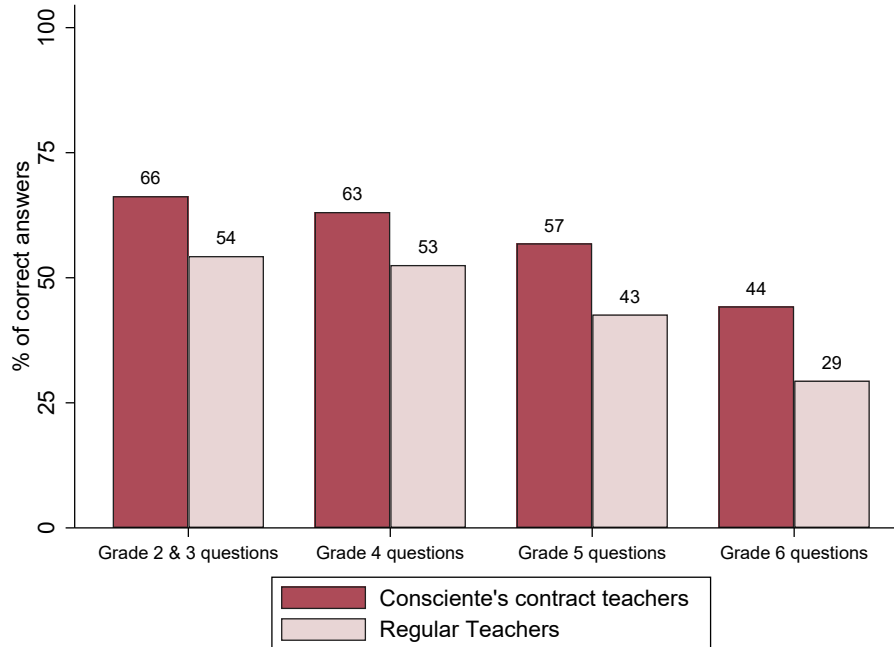
Figure 11: Content knowledge of regular teachers in Morazán (N=231) and Consciente's contract teachers (N=40).
Sources: Brunetti et al. (2019)

they teach. Towards the end of the evaluation we therefore conducted a representative survey among regular teachers of third to sixth graders, as well as Consciente's contract teachers hired for CAL-IMPACT.

The main results illustrated in Figure 11 show a disturbing pattern. *First*, the overall performance in the math assessment covering the curriculum from grades two to six was alarmingly bad. On average, the teachers answered only about half of the questions correctly. *Second*, regular teachers performed even worse than the contract teachers hired by Consciente, scoring correctly on a shocking 29% of sixth grade questions and and 54% of second & third grade questions. These findings suggest, that providing additional math lessons may not only be ineffective when a NGO hires contract teachers, but also when the government asks regular staff to expand school time to the afternoon. Considering the far-reaching implications of these results, the University of Bern together with Consciente launched a follow-up project that evaluates in-service training for teachers (see Brunetti et al., 2019).

Finally, we can compare the estimated impact magnitudes for Consciente's computer-assisted learning intervention to findings from other evaluations on technology-based learning. To do so, we linearly re-scale our impact estimates for the standardized scores to a hypothetical intervention with 100 hours additional CAL-lessons, and add it to Figure 2 discussed in Section 3. Of course, comparing impact estimates across research designs and contexts always comes with caveats, nonetheless it can be informative.

As Figure 12 shows, the impact estimates for CAL+TEACHER are slightly weaker but of comparable magnitude as the average impact reported in earlier evaluations on computer-assisted learning.

19
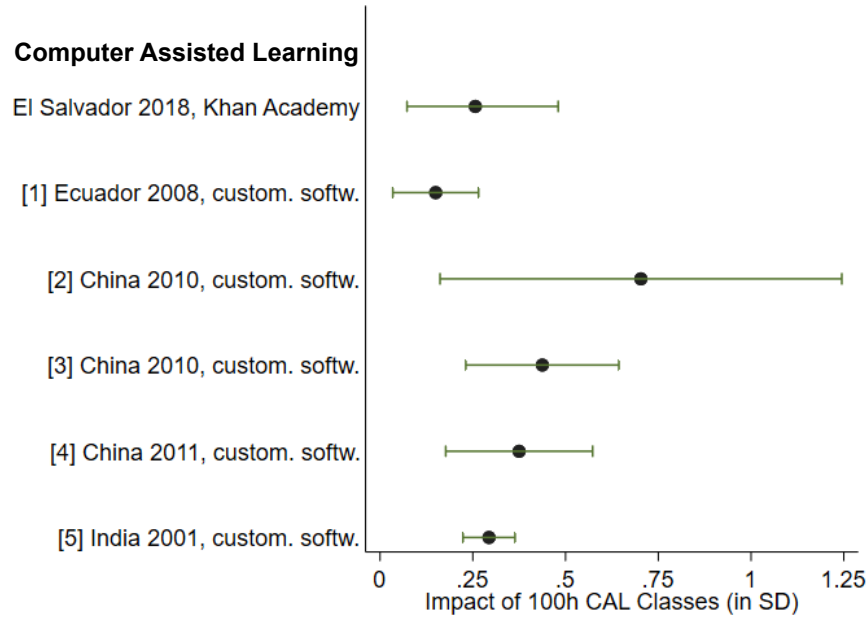
Figure 12: Comparing the effectiveness of CAL+TEACHER to other CAL-interventions.
Source El Salvador: The point estimate resemble the average of the point estimates for CAL+TEACHER in Column (4) of Tables A.1 and A.2. The lower bound confidence interval is based on the estimate's 90%-lower bound in Table A.1, while the upper bound confidence interval is based on the estimates' 90%-upper bound in Table A.2. Point estimates and bounds were linearly rescaled to 100 hours of CAL-lessons.
Other sources: [1] Carrillo, Onofa and Ponce (2011), [2] Lai et al. (2015), [3] Yang et al. (2013), [4] Mo et al. (2014) [5] Banerjee et al. (2007)

Considering that the interventions in China, India and Ecuador were built around a specifically customized learning software, Consciente's intervention based on an off-the-shelf software performed respectably. Having said that, the illustration also suggests that there is some room for improving the implementation of CAL-IMPACT in order to tap its full potential. In this respect, Section 6.2 discussed the (strongly) mitigating effect of student absenteeism, which also shows in this cross-study comparison. For instance, the CAL-courses in China were mandatory, achieved almost full attendance and, as a consequence, higher impacts. As the qualitative feedback from stakeholders surfaced, student absenteeism is not the only challenge that Consciente's staff faced during the program implementation. The next section discusses the perceptions of the major agents involved, including specific suggestions, how the effectiveness of CAL-IMPACT may be improved.

6.4 Qualitative feedback from stakeholders

As a complement to the impact evaluation, extensive qualitative feedback was collected from the different actors involved in the pilot project. These opinions expressed by teachers, school principals, children, the Ministry of Education and the local management are a valuable basis for the further project development. They deepen the understanding of the mechanisms behind the achieved impact and help to appreciate in detail which aspects of the project worked best, and where further

improvement might be needed.

General perceptions Overall, the qualitative evaluation gave the impression that the key stakeholders are highly satisfied with CAL-IMPACT.

Most notably, Consciente received valuable support from *local authorities* due to the program's excellent alignment with the official basic education strategy. CAL-IMPACT was co-coordinated with the Ministry of Education, which not only granted formal access to all schools but also made substantial in-kind contributions. The local director of the Education Ministry characterizes CAL-IMPACT as an essential step towards more participative and modern teaching and wishes to expand it to all schools in Morazán.

Similarly, most *school principals* not only embraced the project with enthusiasm but also showed a high degree of cooperation and initiative. In general, schools were particularly delighted about the improvement in their IT infrastructure and the resulting opportunity for students to acquire better math and also basic computer skills. Correspondingly, all participating schools wished the project to continue after the pilot phase and numerous additional schools applied for participation.

Concerning our own staff, *teachers hired by Consciente* showed a strong intrinsic motivation related to the project, perceiving it as an opportunity to promote social change through better teaching. When asked, many asserted that CAL-IMPACT changed their perceptions about education. One of the teachers put it like this: "I have now understood something essential: that learning does not have to be boring." We also distributed them a questionnaire, that inquired about different aspects of the CAL-IMPACT program. Figure 13 plots the results, which reinforce their favorable feedback from interviews and conversations.
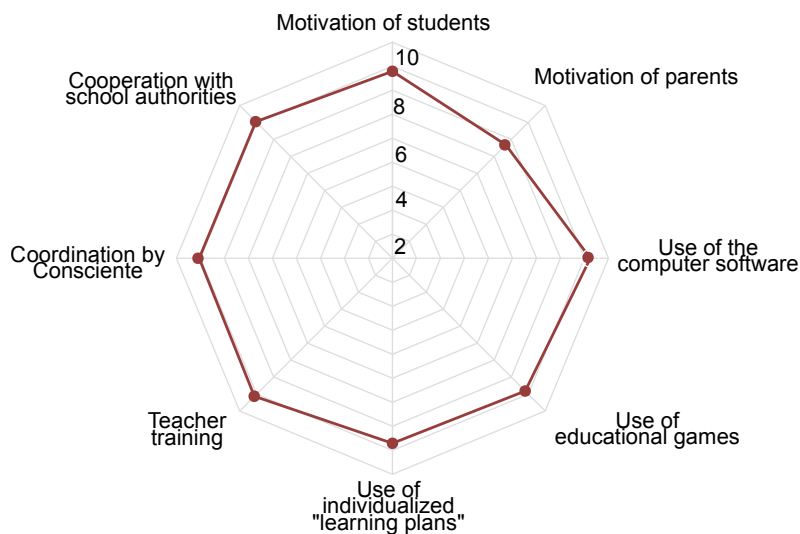

Figure 13: Perceptions of Consciente's contract teachers about different aspects of the project as expressed in an online-survey. The scale runs from 1=terrible to 10=excellent, N=40.
Source: https://www.consciente.ch/surveys/docentes2018/

As to *parents*, a majority strongly approved of the project and appreciated its innovative techno-

logical component. Some mothers even participated by volunteering for lunch preparation. Statements from parents suggest that the project induced many children to develop a greater motivation for math and lose their fear of this "hard subject".

Most importantly, despite a widespread dislike of math, CAL-IMPACT was also popular among *children*. Other than in traditional lecturing classes, the project allowed them to become the main actor rather than a passive listener in the classroom. The use of the interactive computer software and the educational games was a completely new experience for students, which many described as "fun" or "exciting".

Problems and suggestions for improvement   Despite the overwhelmingly positive feedback, the involved actors also pointed out how they think the project could be improved. The following problems emerged as the main weak spots of CAL-IMPACT, which should be addressed in a next project phase:

- **Assignment of content with *Khan Academy*:** Unlike other educational computer programs (which are not available in Spanish), *Khan Academy* does not automatically tailor instruction to students' learning levels. While teachers were instructed to diagnose students' learning levels and assign appropriate contents respectively, they did not always fully succeed in doing so. In particular, students from higher grades where often found to be working on materials that were too advanced for their actual competency levels (e.g. on grade 4 or 5 materials when they still had problems with grade 2 or 3 materials).
→ It is very important to further instruct teachers on how to assign contents based on students' actual learning levels rather than their grade levels.

- **Student absenteeism:** A central operational concern that has been repeatedly expressed and also shows in the monitoring data relates to student absenteeism. Parents tend to perceive the regular (and graded) morning classes as mandatory, and the afternoon classes as optional. Moreover, many parents want their children to help out in the family business during the afternoon or feel uneasy about them coming home late out of security concerns or problems with public transport. According to the Consciente teachers, a key challenge was the lack of real parental commitment with their children's learning.
→ Student absenteeism might be reduced through stronger involvement of parents, greater co-responsibility on the part of the schools, or by grading student performance in the additional classes.

- **Reading deficiencies:** While most teachers perceived the CAL classes as being more successful than the traditional classes, some also reported difficulties related to the use of software. For example, children could get distracted with other features of the computer such as cameras or calculators, although many distractive functions – like internet and games – were disabled. Moreover, some students showed severe reading difficulties. Slow readers were

often too impatient to really understand the text exercises, hence many children reverted to guessing the answers instead of solving the problem.

$\rightarrow$ To improve the impact of the project, reading deficiencies should be addressed too. More generally, teachers should play an active role in the classroom helping children when they are stuck and preventing them from getting distracted.

- **Inadequate infrastructure:** In some schools, the project was confronted with problems regarding infrastructure, e.g. when too few classrooms were available or the electrical installations proved to be inadequate. As a result, in some cases, two classes had to be taught in the same room or our teachers had to arrive much earlier to charge the laptops.

$\rightarrow$ In subsequent phases, the project could benefit from a careful prior check of the infrastructure capacities of participating schools.

- **Collaboration with regular school staff:** While generally high, the cooperativeness of the regular faculty varied substantially between different schools. Strongly committed principals stand in contrast with isolated cases of unsatisfactory collaboration, e.g. when principals were reluctant to make existing IT infrastructure available for the project or to contribute to school meals. Some principals also complained about other schools receiving more computers or larger subsidies for meals.

$\rightarrow$ If resources are limited, the effectiveness of the project can be increased by giving priority to particularly cooperative schools.

Overall, qualitative feedback shows that the CAL-IMPACT pilot project was perceived very favorably by local agents, raising many hopes in the department of Morazán. However, it is also important that experiences of the pilot phase will feed into the continuous improvement of future project phases and help to iron out the main deficiencies.

7 Conclusion

Based on an experimental design, this evaluation rigorously examined a basic education intervention implemented by Consciente in Morazán, El Salvador. The evaluation's core feature is a systematic comparison between the program's three intervention arms, which allows to gather important evidence on their relative cost-effectiveness. We thus circumvent problematic comparisons across contexts and research designs that typically rely on heroic assumptions regarding external validity.

The NGO was interested to learn, whether it makes sense to equip schools with computers and software to conduct computer-assisted learning courses. While several international evaluations suggest, that computer-assisted learning can improve the math ability of the participants, it was yet to be shown that computer-assisted learning outperforms additional math lessons instructed by teachers. Given the results of this evaluation, we learned that at least in the setting of primary schooling in El Salvador, the additional effort and money to setup computer-assisted learning courses pays off. The more traditional math lessons offered by Consciente proved to be either

ineffective (lower bound of our estimates) or had a comparatively small impact (upper bound of our estimates) on learning outcomes of school children. The evaluation reveals that the use of computers not only increased the overall impact of the additional math lessons, but was also more cost-effective. A follow-up survey that we conducted towards the end of the evaluation points to one potentially alarming explanation: Even teachers struggle with the basic math curriculum that they teach; this is both true for the contract teachers that Consciente hired, but even more so for the average teachers working in Morazán's primary schools. Hence, computers with learning software may serve as an effective substitute for the lack of adequately qualified teaching staff. Considering the far-reaching implications of these results, the University of Bern together with Consciente launched a follow-up project that evaluates in-service training for teachers (see Brunetti et al., 2019).

We further compared two different versions of computer-assisted learning lessons in order to identify its complementarities with pedagogically trained teaching staff. While teachers instructing children in computer-assisted learning lessons outperformed supervisors without educational background, the differences in learning gains were statistically insignificant. Moreover, the slightly higher impact of teachers instructing computer-assisted learning lessons compared to supervisors was fully compensated by the proportional increase in program costs due the teachers' higher wage bill. Hence, in terms of cost-effectiveness the two computer-assisted learning interventions perform about equally well: Teachers achieve a slightly stronger impact, but this comes at the expense of proportionally higher labor costs.

Somewhat surprisingly, we found that the presence of Consciente at primary schools even boosted learning among children not directly targeted by the program. From a policy perspective, these type of positive spillovers are of course an attractive feature as they increase the aggregate effect of the intervention on learning outcomes. In the present context, the documented spillover effects seem particularly relevant (and welcome) considering the national policy to collaborate with NGOs to expand school time in the afternoon. The fact that our study finds considerable spillovers are a strong argument in favor of scaling up one of the computer-assisted learning interventions. Since the channels at work could not be identified in this evaluation, the detailed mechanics seem to warrant a closer look. In this respect, our approach to work with two control groups, one within program schools and one in control schools, is well suited to develop further insights on this matter.

These findings demonstrate the benefits of an evidence-based approach for NGOs and policy-makers alike. Rigorous impact evaluations during the pilot phase of the implementation allow us to systematically learn about the program's impact and mechanics. Hence, they are a powerful tool that can help decision-makers to form a well-grounded view on whether public money would be wisely spent on the scale up of the evaluated program. Considering that a priori it is often unclear what works in a given context and what is ineffective, rigorous evaluations are an essential ingredient to weed out ineffective policies and trigger an evolution towards more effective development cooperation.

In summary, we derive *five main conclusions* that may prove valuable for the Salvadorian Ministry of Education and Consciente's future work in the field of basic education:

i.) Computer-assisted learning methods offer a new – and for the children motivating – stimulus that demonstrably improves their basic math skills.

ii.) From a cost-effectiveness perspective, hiring teachers or (less-qualified) supervisors to teach computer-assisted lessons is about equivalent.

iii.) Simply providing additional math lessons (i.e. more of the same) may be a rather ineffective way to improve learning outcomes in math.

iv.) Indirect (positive) effects of the NGOs presence at schools – even on those children not directly targeted by the program – are a strong additional argument in favor of expanding Consciente's initiative to new schools. Since this study could not reliably pin down the underlying channels, this seems to warrant a follow-up analysis.

v.) Deficits in the content knowledge of teachers are a significant source of inefficiencies in the basic math education at Morazán's primary schools. Targeted measures to mitigate these deficits, such as specifically tailored teacher training programs, may be an effective strategy towards sustainable long-term improvements in the quality of basic education.

A Technical appendix: Results

A.1 Benchmark estimations: The effect of CAL-Impact on learning outcomes

The benchmark models aim at obtaining *Intention-to-Treat (ITT) estimates*, based on the following estimation equation:

$$Y_{ics}^{EL} = \alpha + \beta_1 T1_{cs} + \beta_2 T2_{cs} + \beta_3 T3_{cs} + \delta Y_{ics}^{BL} + \gamma_x X_{ics} + \lambda_v V_{cs} + \mu_s + \epsilon_{ics}$$

where, Y_{ics}^{EL} is the endline math score of student i in class c and school s. $T1$, $T2$ and $T3$ are binary indicators for treatment 1 (i.e. CAL+TEACHER), treatment 2 (i.e. CAL+SUPERVISOR) and treatment 3 (i.e. TEACHER). Y_{ics}^{BL} is the baseline math score, X_{ics} represents individual-level controls collected at baseline, and V_{cs} covers classroom-level controls. μ_c are school fixed effects in models using within school control units, and stratum fixed effects in models based on pure control units from control schools. We do not use school fixed effects with control schools, because mostly we collected only data on one class per control school.

The results for within-school comparisons are reported in Table A.1, while Table A.2 displays estimates using pure control units. The two tables are uniformly structured, and show three set of results using different math outcome measures: percent of correct answers (columns 1-2), standardized test score (columns 3-4) and math ability based on Item Response Theory (columns 5-6).

For each outcome, the Tables include one model with a minimal set of control variables and another including all available covariates. Results in Table A.1 may be interpreted as lower bound and results in Table A.2 as upper bound.

The results from the lower bound estimates in Table A.1 may be summarized as follows: Students assigned to CAL+TEACHER perform significantly better than students from the control group; this result is robust across specifications. For instance, columns 3 & 4 suggest that CAL-lessons instructed by teachers raised the participants' standardized math score by 0.11σ. The effect of CAL+SUPERVISOR is slightly lower (i.e. 0.09σ) and depending on the math outcome just above or below the 10% significance-level. The effect size of the treatment without learning software is close to zero (i.e. 0.03σ), and consistently insignificant. Under the assumption that the effect of CAL+TEACHER and TEACHER is the same, the probability to obtain the observed difference in estimates is always around 10%.

In the upper bound estimations shown in Table A.2, we find a positive and highly significant effect for all treatments. However, point estimates of the CAL interventions (about 0.25σ in columns 3 & 4) are still 32% higher than those of the intervention without learning software (about 0.17σ).

Table A.1: Lower bound ITT-estimates: based on control classes from program schools

	Percent Correct		Std. Scores		IRT-Scores	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatments						
T1: CAL-lessons with teacher	1.826**	1.727**	0.118**	0.112**	0.093*	0.087*
	(0.876)	(0.841)	(0.057)	(0.055)	(0.049)	(0.047)
T2: CAL-lessons with supervisor	1.358	1.229	0.097*	0.091*	0.069	0.062
	(0.853)	(0.825)	(0.055)	(0.053)	(0.048)	(0.047)
T3: Lessons with teachers	0.432	0.350	0.038	0.033	0.027	0.023
	(0.861)	(0.814)	(0.056)	(0.053)	(0.049)	(0.047)
Sociodemographics						
Female students		0.766**		0.055**		0.045**
		(0.369)		(0.024)		(0.021)
Age (std. by grade level)		−1.126***		−0.072***		−0.063***
		(0.213)		(0.014)		(0.013)
Household size		−0.116		−0.008		−0.009*
		(0.086)		(0.006)		(0.005)
Household assets		0.407		0.027		0.004
		(1.010)		(0.067)		(0.058)
Class room variables						
Class size (log)		0.091		−0.018		0.005
		(1.460)		(0.094)		(0.085)
Female teacher		2.390***		0.152***		0.131**
		(0.878)		(0.055)		(0.051)
Teacher math score		2.867		0.168		0.154
		(2.300)		(0.149)		(0.131)
Baseline performance						
Baseline math score	0.765***	0.753***	0.775***	0.764***	0.845***	0.833***
	(0.016)	(0.016)	(0.014)	(0.014)	(0.017)	(0.017)
R^2	0.66	0.67	0.66	0.67	0.70	0.70
Observations	2539	2539	2539	2539	2539	2539
$\beta_{T4} := \beta_{T3} - \beta_{T1} = 0$	1.394	1.377	0.081	0.079	0.066	0.065
p-value ($\beta_{T4}=0$)	0.081	0.069	0.127	0.116	0.143	0.129
$\beta_{T5} := \beta_{T3} - \beta_{T2} = 0$	0.468	0.498	0.021	0.021	0.023	0.025
p-value ($\beta_{T5}=0$)	0.567	0.537	0.696	0.692	0.605	0.580
School & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.

Table A.2: Upper bound ITT-estimates: based on pure control classes from control schools

	Percent Correct		Std. Scores		IRT-Scores	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatments						
T1: CAL-lessons with teacher	3.968***	3.950***	0.254***	0.256***	0.215***	0.214***
	(0.898)	(0.922)	(0.059)	(0.060)	(0.050)	(0.052)
T2: CAL-lessons with supervisor	3.483***	3.555***	0.233***	0.241***	0.192***	0.197***
	(0.851)	(0.850)	(0.055)	(0.055)	(0.049)	(0.049)
T3: Lessons with teachers	2.348***	2.455***	0.158***	0.168***	0.134**	0.141***
	(0.895)	(0.938)	(0.058)	(0.061)	(0.052)	(0.053)
Sociodemographics						
Female students		0.917**		0.065***		0.058***
		(0.363)		(0.024)		(0.021)
Age (std. by grade level)		−1.204***		−0.076***		−0.066***
		(0.231)		(0.015)		(0.015)
Household size		−0.035		−0.003		−0.004
		(0.089)		(0.006)		(0.005)
Household assets		−1.475		−0.089		−0.087
		(0.968)		(0.064)		(0.058)
Class room variables						
Class size (log)		−0.561		−0.061		−0.049
		(1.159)		(0.075)		(0.065)
Female teacher		1.235		0.068		0.069
		(0.830)		(0.054)		(0.048)
Teacher math score		4.826**		0.307**		0.267**
		(2.084)		(0.132)		(0.119)
Baseline performance						
Baseline math score	0.773***	0.763***	0.782***	0.773***	0.858***	0.847***
	(0.017)	(0.017)	(0.015)	(0.015)	(0.016)	(0.016)
R^2	0.65	0.66	0.66	0.67	0.69	0.70
Observations	2565	2565	2565	2565	2565	2565
$\beta_{T4} := \beta_{T3} - \beta_{T1} = 0$	1.62	1.495	0.096	0.087	0.081	0.073
p-value ($\beta_{T4}=0$)	0.093	0.111	0.126	0.152	0.135	0.161
$\beta_{T5} := \beta_{T3} - \beta_{T2} = 0$	0.486	0.394	0.021	0.014	0.024	0.017
p-value ($\beta_{T5}=0$)	0.597	0.663	0.717	0.808	0.641	0.728
Stratum & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.

A.2 Spillover effects and potential channels

In Table A.3 we compare the 40 control classes from the program schools with the 40 classes from the control schools. Based on this sample we estimate the following equation:

$$Y_{ics}^{EL} = \alpha + \beta_1 C_s + \delta Y_{ics}^{BL} + \gamma_x X_{ics} + \lambda_v V_{cs} + \mu_s + \epsilon_{ics}$$

where C_s is a binary indicator for control schools. All other variables are defined as in the benchmark estimation in Appendix A.1. The results show that students from control schools score significantly lower than students from control classes within treatment schools. Hence, this makes positive spillovers from treatment classes to control classes in the same school very likely. The size of the spillover effect is about of similar magnitude as the lower-bound impact of CAL-lessons instructed by teachers derived from the within school comparison.

Table A.3: Estimating spillover effects: control classes from program schools vs. pure control classes from control schools

	Percent Correct		Std. Scores		IRT-Scores	
	(1)	(2)	(3)	(4)	(5)	(6)
Pure Control School	−1.95**	−2.00**	−0.12**	−0.12**	−0.11**	−0.11**
	(0.95)	(0.85)	(0.06)	(0.06)	(0.05)	(0.05)
Sociodemographics						
Female students		0.61		0.05		0.04
		(0.50)		(0.03)		(0.03)
Age (std. by grade level)		−1.61***		−0.11***		−0.10***
		(0.24)		(0.01)		(0.02)
Household size		−0.14		−0.01		−0.01
		(0.12)		(0.01)		(0.01)
Household assets		1.49		0.10		0.06
		(1.34)		(0.09)		(0.08)
Class room variables						
Class size (log)		2.30		0.14		0.14
		(1.99)		(0.13)		(0.11)
Female teacher		1.25		0.08		0.08
		(1.26)		(0.08)		(0.07)
Teacher math score		4.41*		0.24		0.26*
		(2.44)		(0.16)		(0.15)
Baseline performance						
Baseline math score	0.78***	0.77***	0.80***	0.79***	0.87***	0.86***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
R^2	0.69	0.70	0.68	0.69	0.72	0.74
Observations	1274	1274	1274	1274	1274	1274
Stratum & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered standard errors in parentheses, * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Student attendance in regular classes We can use monitoring data to examine whether the presence of the NGO changed the behavior of students or school staff. Observing such changes would reveal specific channels of spillover effects. In Table A.4 we estimate the program's effect on attendance of students in regular classes. For instance, students' motivation to show up in regular classes might change, if they have to attend additional math classes in the afternoon. Motivation of students from within school control classes might change too. If students from the within school control group observe that peers from treatment classes are offered additional math classes, they might be more (or even less) motivated to show up in regular classes themselves. However, we do not find any evidence supporting the hypothesis that CAL-IMPACT affected students' attendance in regular classes. Attendance rates in regular classes do not differ between treatment and control classes (columns 1–4) nor between the two control groups (columns 5–6).

Table A.4: The effect of Consciente's presence on student attendance in regular classes

Control classes from:	Within school		Control schools		Both: Within & Control	
Dep. var.: Student Attendance in %	(1)	(2)	(3)	4)	(5)	(6)
Treatments						
T1: CAL-lessons with teacher	−1.20	−1.18	−0.91	−0.29		
	(1.37)	(1.36)	(2.09)	(2.09)		
T2: CAL-lessons with supervisor	0.25	0.22	0.46	1.02		
	(1.38)	(1.40)	(1.84)	(1.83)		
T3: Lessons with teachers	0.29	0.30	0.71	1.41		
	(1.32)	(1.37)	(1.84)	(1.89)		
Pure Control School					−0.07	−0.39
					(1.72)	(1.75)
Class room variables						
Class size (log)		0.62		1.02		1.42
		(2.07)		(2.22)		(2.22)
Female teacher		−0.14		−0.68		0.04
		(1.37)		(1.44)		(1.18)
Teacher math score		2.78		6.75**		5.12**
		(2.39)		(3.11)		(2.53)
R^2	0.50	0.50	0.10	0.14	0.10	0.12
Observations	158	158	158	158	198	198
School FE	Yes	Yes	No	No	No	No
Stratum FE	No	No	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: clustered standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.

Cancellations of regular lessons Another possible channel for spillovers might be that the presence of the NGO lowered cancellation rates of regular lessons. But, there might be as well a within school effect on class cancellations due to the program. For instance, teachers know that their students are having additional math lessons and therefore, their motivation to conduct the regular lessons might change. On the one hand, the motivation might increase if regular teachers perceive the additional lessons as competitors of their regular lessons. On the other hand, the motivation might decrease if regular teachers think that students get enough math lessons elsewhere and that they might as well prioritize other subjects. Table A.5 examines the effect of CAL-IMPACT on the cancellation rate of regular lessons. We find no evidence for an effect on the number of regular lessons canceled, neither when comparing treatment classes to the within school control group (columns 1–4) nor when comparing the control groups from program and control schools (columns 5 & 6).

Table A.5: The effect of Consciente's presence on the cancellation of regular classes

Control classes from:	Within school		Control schools		Both: Within & Control	
Dep. var.: Classes cancelled in %	(1)	(2)	(3)	4)	(5)	(6)
Treatments						
T1: CAL-lessons with teacher	−3.49	−3.30	2.02	2.52		
	(2.33)	(2.36)	(4.75)	(5.06)		
T2: CAL-lessons with supervisor	0.83	1.11	6.87	7.70		
	(2.80)	(2.94)	(4.67)	(5.10)		
T3: Lessons with teachers	−1.51	−1.22	4.93	5.60		
	(2.33)	(2.33)	(5.08)	(5.31)		
Pure Control School					−4.85	−4.87
					(4.62)	(4.90)
Class room variables						
Class size (log)		−0.21		−2.93		−1.08
		(5.70)		(6.17)		(5.54)
Female teacher		−3.88*		−3.99		−2.07
		(2.03)		(4.28)		(3.64)
Teacher math score		−1.59		−4.78		−6.28
		(4.13)		(8.14)		(6.58)
R^2	0.71	0.72	0.08	0.10	0.08	0.08
Observations	158	158	158	158	198	198
School FE	Yes	Yes	No	No	No	No
Stratum FE	No	No	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: clustered standard errors in parentheses, * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

A.3 IV-estimates: Hypothetical scenario with full attendance

The benchmark results in Appendix A.1 are ITT-estimates that do not account for the actual attendance rate of students in the additional math lessons. The instrumental variable (IV) approach estimates the impact of the three intervention under the assumption of full attendance.

The validity of these results depend on two additional assumptions. *First*, the treatment effect needs to be homogenous across students. *Second*, the functional form between attendance and treatment effects should be linear. In our case the functional form appears to be quadratic with the squared attendance entering positively. Hence, our IV-estimates tend to be downward biased.

The IV-models are estimated in two-stage least square estimations, with the first-stage estimation being specified as:

$$Att_{ics}^{T=t} = \alpha + \beta_1 T1_{cs} + T2_{cs} + \beta_3 T3_{cs} + \rho Y_{ics}^{BL} + \phi_x X_{ics} + \tau_v V_{cs} + \pi_s + \nu_{ics} \quad for \quad t \in [1,2,3]$$

where $Att_{ics}^{T=t}$ is the attendance rate of one of the three treatments t and is always zero for all students that were not assigned to treatment t. All other variables are as defined in the estimation equation in Appendix A.1. In the second stage, we estimate basically the same equation as in Appendix A.1, but replace the binary treatment indicators with the predicted attendance rates $\widehat{Att}_{ics}^{T=t}$ from stage 1:

$$Y_{ics}^{EL} = \alpha + \beta_1 \widehat{Att}_{ics}^{T=1} + \beta_2 \widehat{Att}_{ics}^{T=2} + \beta_3 \widehat{Att}_{ics}^{T=3} + \delta Y_{ics}^{BL} + \gamma_x X_{ics} + \lambda_v V_{cs} + \mu_s + \epsilon_{ics}$$

We estimate the IV-models based on two different control groups: Table A.6 uses within program school control classes, while Table A.7 is based on pure control classes from control schools. Again, three different math outcomes are used: percent of correct answers (columns 1-2), standardized test score (columns 3-4) and math ability based on Item Response Theory (columns 5-6).

Under the assumption of full attendance, effects increase because of the positive correlation between attendance and math outcomes. The results from the lower bound estimates in Table A.6 may be summarized as follows: The impact of CAL-classes instructed by teachers is again the highest and significant at the 95%-level for all models. For instance, columns 3 & 4 suggest that CAL-lessons instructed by teachers raised the participants' standardized math score by 0.18σ. The effect of CAL+SUPERVISOR is slightly lower (i.e. 0.16σ) and depending on the math outcome just above or below the 10% significance-level. The effect size of the treatment without learning software is close to zero (i.e. 0.05σ), and consistently insignificant. The probability that the effect of CAL+TEACHER is higher than the effect of TEACHER meanders around the 10% significance threshold.

In the upper bound estimations shown in Table A.7, we find a positive and highly significant effect for all treatments. However, point estimates of the CAL interventions (about 0.42σ) are still considerably higher than those of the intervention without learning software (about 0.30σ).

Table A.6: Lower bound IV-estimates: based on control classes from program schools

	Percent Correct		Std. Scores		IRT-Scores	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatments						
T1: CAL-lessons with teacher	2.916**	2.764**	0.189**	0.179**	0.150*	0.141*
	(1.351)	(1.302)	(0.089)	(0.085)	(0.077)	(0.075)
T2: CAL-lessons with supervisor	2.369	2.116	0.170*	0.156*	0.123	0.109
	(1.473)	(1.398)	(0.094)	(0.089)	(0.085)	(0.081)
T3: Lessons with teachers	0.688	0.558	0.061	0.053	0.044	0.037
	(1.410)	(1.334)	(0.091)	(0.087)	(0.082)	(0.077)
Sociodemographics						
Female students		0.734**		0.053**		0.044**
		(0.367)		(0.024)		(0.022)
Age (std. by grade level)		−1.067***		−0.068***		−0.061***
		(0.209)		(0.013)		(0.013)
Household size		−0.121		−0.008		−0.009*
		(0.084)		(0.005)		(0.005)
Household assets		0.384		0.025		0.003
		(0.997)		(0.066)		(0.058)
Class room variables						
Class size (log)		0.361		0.001		0.019
		(1.425)		(0.091)		(0.084)
Female teacher		2.365***		0.151***		0.132***
		(0.857)		(0.054)		(0.051)
Teacher math score		3.039		0.181		0.166
		(2.245)		(0.145)		(0.131)
Baseline performance						
Baseline math score	0.760***	0.749***	0.770***	0.759***	0.840***	0.828***
	(0.015)	(0.016)	(0.014)	(0.014)	(0.017)	(0.017)
R^2	0.67	0.68	0.67	0.68	0.70	0.71
Observations	2539	2539	2539	2539	2539	2539
Kleibergen-Paap F-statistic	354.53	380.54	353.74	379.83	354.16	380.24
$\beta_{T4} := \beta_{T3} - \beta_{T1} = 0$	2.228	2.206	0.128	0.126	0.106	0.105
p-value ($\beta_{T4}=0$)	0.078	0.067	0.125	0.116	0.142	0.129
School & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered standard errors in parentheses, * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table A.7: Upper bound IV-estimates: based on pure control classes from control schools

	Percent Correct		Std. Scores		IRT-Scores	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatments						
T1: CAL-lessons with teacher	6.503***	6.564***	0.417***	0.425***	0.358***	0.361***
	(1.387)	(1.452)	(0.091)	(0.094)	(0.079)	(0.083)
T2: CAL-lessons with supervisor	6.275***	6.492***	0.419***	0.440***	0.351***	0.364***
	(1.504)	(1.479)	(0.098)	(0.096)	(0.088)	(0.088)
T3: Lessons with teachers	4.054***	4.367***	0.273***	0.298***	0.236***	0.254***
	(1.481)	(1.543)	(0.097)	(0.100)	(0.087)	(0.089)
Sociodemographics						
Female students		0.777**		0.056**		0.051**
		(0.362)		(0.024)		(0.021)
Age (std. by grade level)		−1.041***		−0.065***		−0.058***
		(0.230)		(0.015)		(0.015)
Household size		−0.054		−0.005		−0.005
		(0.087)		(0.006)		(0.005)
Household assets		−1.440		−0.087		−0.086
		(0.961)		(0.064)		(0.059)
Class room variables						
Class size (log)		−0.297		−0.043		−0.035
		(1.156)		(0.075)		(0.066)
Female teacher		1.176		0.064		0.067
		(0.802)		(0.053)		(0.047)
Teacher math score		5.301***		0.338***		0.298***
		(1.967)		(0.124)		(0.114)
Baseline performance						
Baseline math score	0.760***	0.751***	0.769***	0.760***	0.845***	0.834***
	(0.017)	(0.017)	(0.015)	(0.015)	(0.016)	(0.016)
R^2	0.66	0.67	0.67	0.68	0.70	0.71
Observations	2565	2565	2565	2565	2565	2565
Kleibergen-Paap F-statistic	303.31	230.57	302.42	229.76	303.1	230.75
$\beta_{T4} := \beta_{T3} - \beta_{T1} = 0$	2.449	2.197	0.144	0.126	0.123	0.107
p-value ($\beta_{T4}=0$)	0.102	0.132	0.141	0.183	0.151	0.194
Stratum & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.

B Technical appendix: Methods & data

B.1 Balance at baseline

Table B.1: Balance at baseline

Panel A: Math Scores (N=3532)	(1) V1: CAL w. Teacher	(2) V2: CAL w. Supervisor	(3) V3: Math w. Teacher	(4) Within School Controls	(5) Pure Control Classes	(6) p-value
%-share correct answers	31.95 (2.06)	33.47 (1.90)	30.33 (1.79)	32.64 (1.31)	30.80 (2.00)	0.44
Standardized math score	−0.04 (0.11)	0.00 (0.11)	−0.15 (0.09)	0.00 (0.07)	−0.14 (0.11)	0.39
IRT math score	−0.08 (0.16)	0.01 (0.14)	−0.16 (0.14)	−0.08 (0.10)	−0.17 (0.15)	0.71
Panel B: Sociodemographics (N=3532)						
Female student	0.51 (0.04)	0.48 (0.04)	0.47 (0.04)	0.47 (0.03)	0.46 (0.04)	0.79
Student age	10.47 (0.28)	10.59 (0.26)	10.30 (0.27)	10.37 (0.19)	10.43 (0.25)	0.85
Household size	5.57 (0.12)	5.61 (0.12)	5.56 (0.13)	5.56 (0.08)	5.50 (0.12)	0.92
Household assets index	0.54 (0.02)	0.55 (0.02)	0.55 (0.02)	0.56 (0.02)	0.56 (0.02)	0.88
Panel C: Class room variables (N=198)						
Class size	18.54 (0.77)	20.05 (1.08)	18.93 (0.76)	18.45 (1.09)	18.20 (1.86)	0.46
Absence rate at baseline (%)	8.85 (1.64)	7.63 (2.17)	7.66 (1.77)	7.94 (1.38)	6.41 (1.75)	0.72
Teacher's math score	0.54 (0.05)	0.56 (0.05)	0.54 (0.04)	0.55 (0.03)	0.63 (0.04)	0.27
Female teacher	0.73 (0.12)	0.80 (0.09)	0.79 (0.10)	0.76 (0.08)	0.54 (0.13)	0.25

Panel D: School variables (N=49)	Treatment Schools	Pure Control Schools	p-value
# classes grade 3–6	5.48 (0.47)	6.25 (0.74)	0.30
Computer lab	0.79 (0.08)	0.75 (0.12)	0.73
Local population density	0.18 (0.01)	0.19 (0.02)	0.62

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 during the baseline survey in February 2018. Column 6 shows the p-value from testing whether the mean is equal across al trement groups, i.e. $H0 \coloneqq$ mean is equal across groups. The household asset index measures what share of the following assets a household owns: Books, electricity, television, washmachine, computer, internet and car. Local density is the municipality's population density measured in 1000 inhabitans per km2. Standard errors are clustered at the class level in Panels A & B, and at the school level in Panel C.

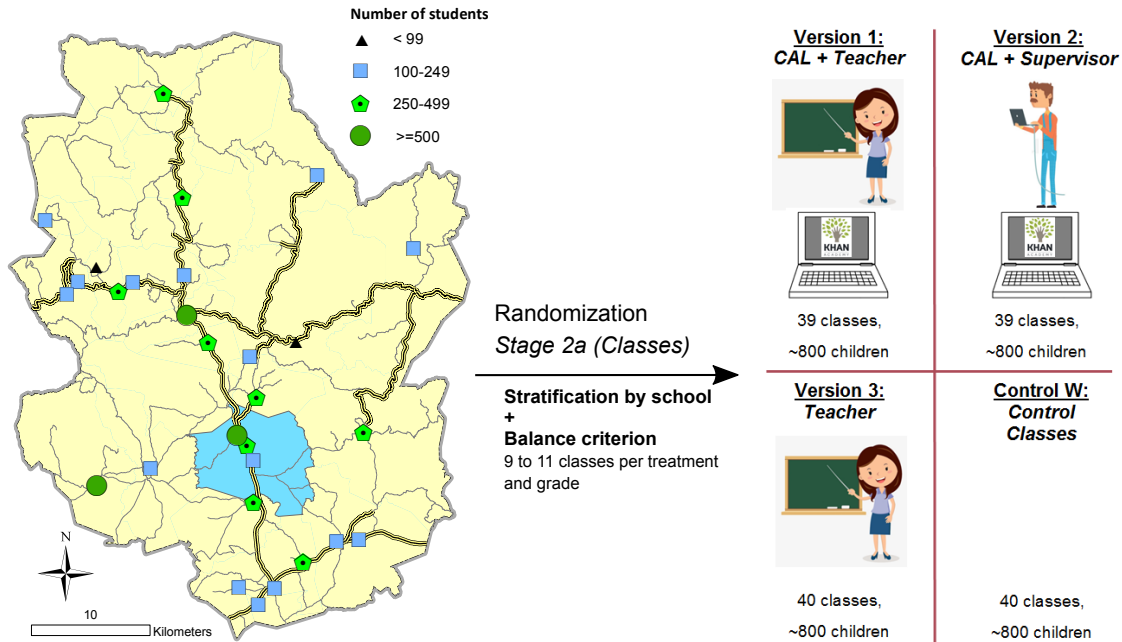## B.2 Sample selection and randomization illustrated
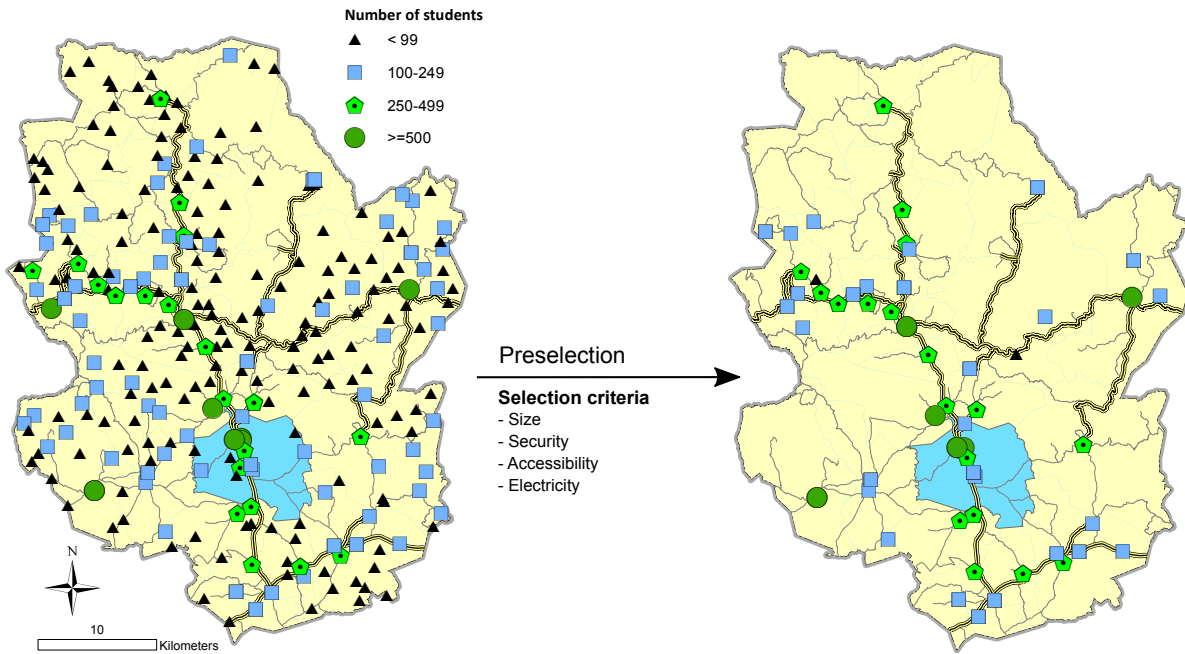
(a) Preselection of eligible primary schools (N=57) from universe of primary schools in Morazán (N=302).
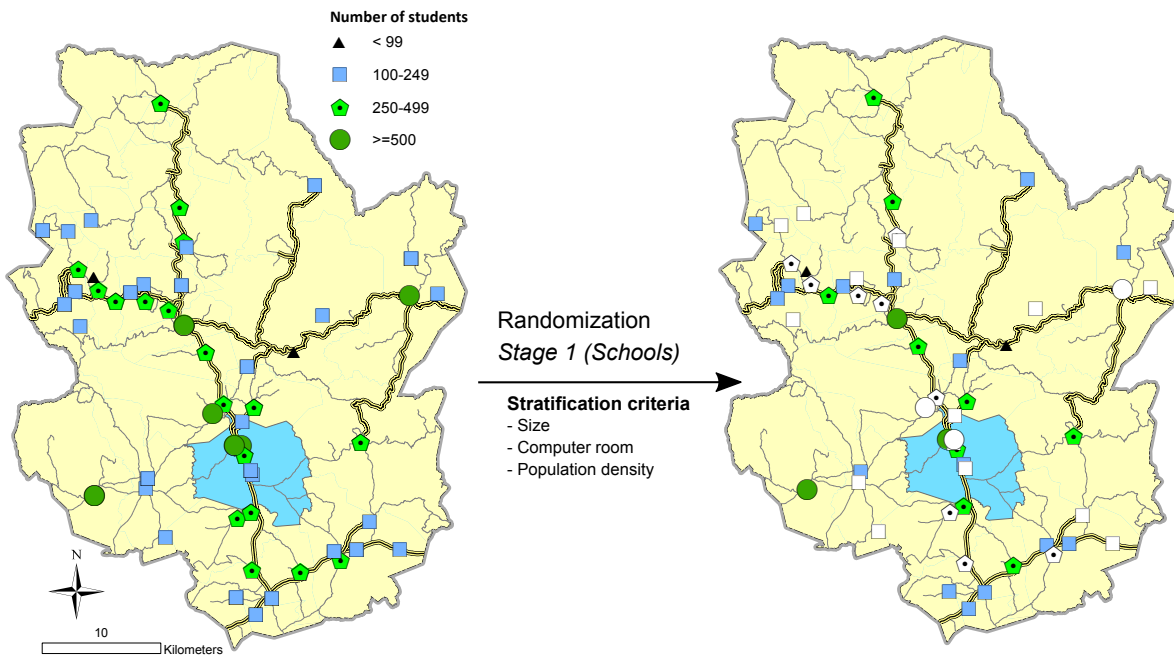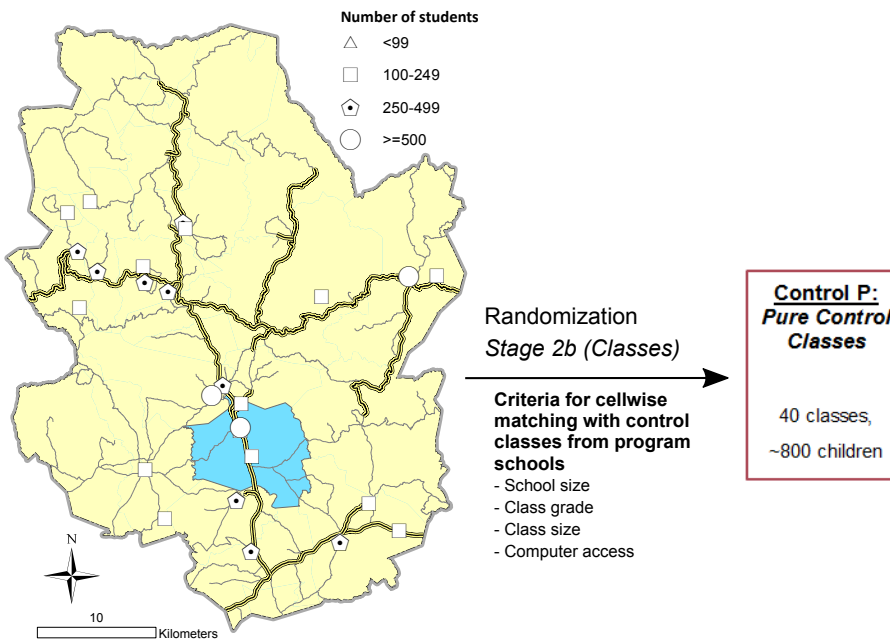

(b) Random assignment of preselected schools (N=57) to the program (N=29, colored) or to pure control schools (N=28, white).

Figure B.1: *(a)* Preselection of schools and *(b)* School level randomization.

(a) Random assignment of classes in the program schools ($N=158$) to the control group ($N=40$) or one of three program versions ($N=118$).

(b) Random cell-wise matching of 40 classes from pure control schools to the 40 control classes from program schools.

Figure B.2: Class level randomization. *(a)* Random assignment of control status and program version 1 to 3 within partner schools. *(b)* Cell-wise random matching of pure control classes from control schools to control classes in partner schools.

B.3 Measuring and converting learning outcomes

To measure measure math skills of third to sixth graders, we conducted two standardized math assessments during the school year 2018 (see Figure 5). These assessments were designed as follows:

1. We summarized the Salvadorian curriculum in math 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 Salvadorian curriculum. These sources are (a.) official text books of El Salvador, (b.) publicly available items from the STAR[12] evaluations in California, (c.) publicly available items from the VERA[13] evaluations in Germany, and (d.) exercises from the Swiss textbook MATHWELT.

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

4. Finally, we designed paper and pencil maths tests using insights from step 3. The 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. Figure B.3 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.
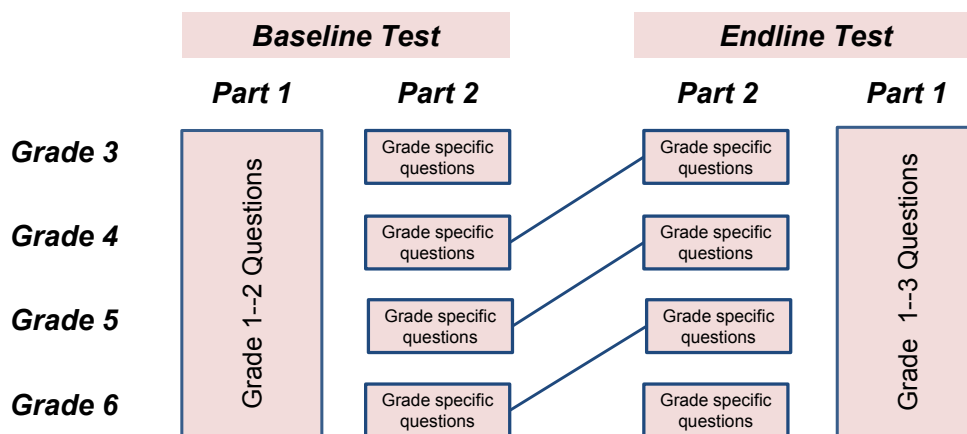

Figure B.3: Stylized illustration of the math assessment design.

Calculating IRT-Scores A particularly nice feature of our math assessments is that they allow 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:

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

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.3 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 a meaningful interpretation, IRT scores were represented as school year equivalents. For this purpose, ability estimates were re-scaled based on between-grade ability differences at the time of the baseline assessment; that means they were divided by the average difference between adjacent grades, which we calculated to be 0.46 in the baseline data. That means, that the average ability difference between third and fourth graders, fourth and fifth graders, and fifth and sixth graders in February 2018 equaled 0.46.

The estimated program effects can then be interpreted as a proportion of children's normal learning 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 do also partly reflect age-based cognitive development, learning at home or spillovers from other subjects (e.g. literacy or science).

C   Appendix: Addressing ethical concerns

Critics of randomized controlled trials often point to ethical concerns related to random assignment and collecting data from control units. For this evaluation all available measures were taken to address such concerns. This section explains, how we proceeded to guarantee that the evaluation meets strict ethical guidelines.

Randomization   The central feature of RCTs is the random assignment of participants to treatment and control units. This step is often criticized as unethical, because the treatment is withheld from certain individuals. In our case, however, random assignment did not lower the number of beneficiaries reached by the NGO, because Consciente faced substantial oversubscription. With oversubscription we mean that Consciente could not reach all eligible schools due to limited financial resources. From an ethical perspective, this has the favorable feature that the number of beneficiaries reached by the NGO is independent of the evaluation's design.

Nature of the intervention   Another ethical reservation is often tied to the intervention itself, especially if people are deceived or not transparently informed about the details. Offering additional math lessons, as in this evaluation, is unambiguously beneficial for all participants, and all activities by Consciente and the University of Bern were closely coordinated with the Ministry of Education.

Data collection   At specifically organized meetings, all relevant stakeholders were informed about the evaluation and the planned data collection. The regional minister of education, as well as the headmasters of all candidate schools received an outline of the evaluation approach and were asked to provide their written consent, before their school was definitely signed up to participate. The parents of the participating children were also invited to information evenings where the project and the evaluation was carefully explained to them. It was also pointed out that participation in the evaluation is voluntary. Finally, on both test days children were informed about the project and the use of the data and were allowed to refuse participation. Consciente and the University of Bern have also committed not to share any data that allows conclusions to be drawn about individuals, classes or specific schools. Furthermore, the socio-demographic survey only asked about standard insensitive characteristics, excluding, for instance, questions about ethnicity, place of birth or religion.

Approval of design by an ethics committee   We provided detailed information about the nature of the intervention and the evaluation design to the Ethics Committee of the *Faculty of Business Administration, Economics and Social Sciences of the University of Bern*. The Ethics Committee concluded that the research design is ethically unproblematic and that the projects complies with the University's ethical standards.

References

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.

Barrera-Osorio, Felipe and Linden Leigh. 2009. "The Use and Misuse of Computers in Education: Evidence from a Randomized Controlled Trial of a Language Arts Program." *Mimeo.*

Brunetti, Aymo, Konstantin Büchel, Martina Jakob, Ben Jann, Christoph Kühnhanss and Daniel Steffen. 2019. "Teacher Content Knowledge in Morazán, El Salvador: Some Stylized Facts." *Ongoing project.*

Carrillo, Paul, Mercedes Onofa and Juan Ponce. 2011. "Information Technology and Student Achievement: Evidence from a Randomized Experiment in Ecuador." *IDB Working Paper Series.*

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.

Cristia, Julián, Pablo Ibarrán, Santiago Cueto, Ana Santiago and Eugenio Severín. 2012. "Technology and Child Development: Evidence from the One Laptop per Child Program." *IZA Discussion Paper.*

de Ayala, R.J. 2009. *The Theory and Practice of Item Response Theory.* New York: Guilford Press.

DIGESTYC, Dirección General de Estadística y Censos El Salvador. 2018. "Encuesta de Hogares de Dirección General de Estadística y Censos 2017 (EHPM)." Online avaliable, URL: `www.digestyc.gob.sv` (25.07.2018).

Günther, Isabel. 2016. "Wirksame Entwicklungshilfe baut auf Fakten." *Die Volkswirtschaft* 3:26–29.

Kudrzycki, Bartlomiej and Isabel Günther. 2018. "Improving Development Policies with Impact Evaluations." *Rural 21* 52:6–8.

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* 47:37–48.

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. 2013. "Elementos para el desarrollo del Modelo Pedagogico del Sistema Educativo Nacional – Escuela Inclusiva de Tiempo Pleno." Ministerio de la Educacion de El Salvador, published online: `https://www.mined.gob.sv/jdownloads/Institucional/modelopedagogico.pdf` (last accessed: 14.01.2018).

Mo, Di, Linxiu Zhang, Jiafu Wang, Weiming Huang, Yao Shi, Matthew Boswell and Scott Rozelle. 2014. "The Persistence of Gains in Learning from Computer Assisted Learning: Evidence from a Randomized Experiment in Rural Schools in Shaanxi Province in China." *REAP Working Paper.*

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

Snilstveit, Birte, Jennifer Stevenson, Daniel Phillips, Martina Vojtkova, Emma Gallagher, Tanja Schmidt, Hannah Jobse, Maisie Geelen, Maria Pastorello and John Eyers. 2015. ""Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle- Income Countries: A Systematic Review"." 3ie Systematic Review 24. London: International Initiative for Impact Evaluation (3ie).

World Bank. 2018. *Learning. To realize eeducation'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.