UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

Inadequate Teacher Content Knowledge and What to Do About It: Evidence from El Salvador

Aymo Brunetti, Konstantin Büchel, Martina Jakob,
Ben Jann, Daniel Steffen

21-14

December, 2021

DISCUSSION PAPERS

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

Inadequate Teacher Content Knowledge and What to Do About It: Evidence from El Salvador*

Aymo Brunetti[a], Konstantin Büchel[a], Martina Jakob[b],
Ben Jann[b], Daniel Steffen[c]

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

December 9, 2021

Abstract

Good teachers are the backbone of a successful education system. Yet, in developing countries, teachers' content knowledge is often inadequate. This study documents that primary school math teachers in the department of Morazán in El Salvador only master 47 percent of the curriculum they teach. In a randomized controlled trial with 175 teachers, we further evaluate a computer-assisted learning (CAL) approach to address this shortcoming. After a five months in-service training combining CAL-based self-studying with monthly workshops, participating teachers outperformed their peers from the control group by 0.29σ, but this effect depreciated by 72 percent within one year. Our simulations show that the program is unlikely to be as cost-effective as CAL interventions directly targeting students.

JEL classification: C93, I20, I21, I28, O15.
Keywords: Education quality, teacher performance, teacher training, student learning, basic math education, computer-assisted learning.

*We are grateful to David Burgherr, Malin Frey, Christoph Kühnhanss, and Amélie Speiser who provided excellent research assistance. The project further benefited from invaluable feedback by Mauricio Romero and participants at the SSES Annual Congress 2021 and the German Development Economics Conference 2021. We further acknowledge generous funding by the IMG Stiftung and the Faculty of Business, Economics and Social Sciences of the University of Bern. Martina Jakob discloses that she serves on a voluntary basis as president of *Consciente – Unterstützungsverein für El Salvador (Schweiz)*. We received IRB approval from the Faculty of Business, Economics and Social Sciences at the University of Bern. A randomized controlled trials registry entry is available at: `https://www.socialscienceregistry.org/trials/4092`.

Contact Details (authors ordered alphabetically):
Brunetti: Univ. of Bern, Dept. of Economics, Schanzeneckstr. 1, CH-3001 Bern, aymo.brunetti@vwi.unibe.ch
Büchel: Univ. of Bern, Dept. of Economics, Schanzeneckstr. 1, CH-3001 Bern, konstantin.buechel@vwi.unibe.ch
Jakob: Univ. of Bern, Inst. of Sociology, Fabrikstr. 8, CH-3012 Bern, martina.jakob@unibe.ch
Jann: Univ. of Bern, Inst. of Sociology, Fabrikstr. 8, CH-3012 Bern, ben.jann@unibe.ch
Steffen: Lucerne University of Applied Sciences and Arts, Suurstoffi 1, CH-6343 Rotkreuz, dani.steffen@hslu.ch

1 Introduction

In light of the persistently low learning levels in many developing countries, it is critical to gain a better understanding of the binding constraints to effective teaching. While various aspects of educational systems such as material inputs, pedagogical practices or teacher incentives have been extensively studied (e.g., Kremer et al., 2013; Glewwe and Muralidharan, 2016), one indispensable precondition to successful instruction has been largely neglected: teachers' content knowledge. Consequently, little is known about the extent to which teachers master the curriculum they have to convey to their students and how to effectively narrow potential knowledge gaps. This paper addresses both questions (see Figure 1). In the first part of the study, we assess the content knowledge of Salvadoran math teachers based on a representative sample of primary school teachers in the department of Morazán. In the second part of the study, we experimentally evaluate an intervention aiming to improve teachers' content knowledge through computer-assisted learning (CAL).

Recent evidence suggests that many primary school teachers may not possess sufficient mastery of the concepts they have to teach. For a sample covering seven sub-Saharan nations, Bold et al. (2017a) asked teachers to mark mock tests and then estimated that only two-thirds of primary school teachers possess minimum proficiency in their subject. In *the first part of our study*, we directly measure teachers' content knowledge through an exam-type assessment with a representative sample of 224 primary school math teachers in the department of Morazán in El Salvador. The average primary school teacher in our sample was able to answer 47 percent of grade two to grade six questions correctly and only 14 percent of the teachers possessed minimum subject proficiency as defined by Bold et al. (2017a). For example, 43 percent of the teachers correctly computed the area of a rectangle, 36 percent were able to add two fractions and a mere 25 percent could retrieve information from a descriptive chart.

Teachers' content knowledge matters. Previous findings suggest that a 1σ increase in teacher content knowledge is associated with a 0.1σ gain in annual student learning (Metzler and Woessmann, 2012; Bau and Das, 2020). Bold et al. (2017a) further document that gaps in the content knowledge of African teachers account for 30 percent of the shortfalls in student learning relative to the curriculum. But how can teachers' subject mastery be improved? Unfortunately, there is little evidence on how to strenghten teacher content knowledge and the impact thereof.[1] A potentially promising approach to teacher professional development is the use of CAL software. A growing strand of literature documents the success of technology-based instruction with students (for reviews see Escueta et al., 2020; Rodriguez-Segura, 2021), and the targeted use of technology may also entail considerable advantages for teacher professional development. For instance, a successful CAL-based training would be relatively easy to replicate and scale, and

[1] In a thorough literature search, we identified 28 experimental and quasi-experimental studies analyzing the impact of teacher professional development in low and middle-income countries. Among those, seven training approaches include a significant content-knowledge component, and only three out of the seven studies actually assess the impact of the treatment on the content-knowledge of teachers (San Antonio et al., 2011; Zhang et al., 2013; Bando and Li, 2014). The three cited studies evaluate trainings that combine pedagogical and content-related elements, and their findings are mixed. Bando and Li (2014) document short term gains on the English skills of teachers and students in Mexico, whereas San Antonio et al. (2011) find a positive impact on math skills of Philippine teachers but not their students. Zhang et al. (2013) study the impact of an intensive three-week training in English for teachers in Chinese schools, and report insignificant treatment effects for both teachers and students.

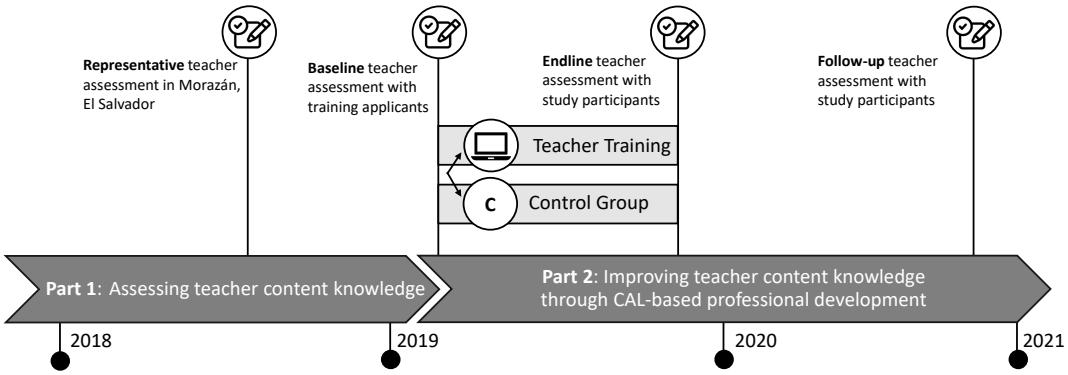


Figure 1: Timeline of the study

hence mitigate concerns about the sensitivity of program effectiveness to details of design and implementation (see Kerwin and Thornton, 2021).

To assess the value of CAL software in teacher professional development, *the second part of this study* presents a randomized controlled trial implemented with 175 math teachers in El Salvador. The treatment consisted of a five-month in-service teacher training program that combined CAL-based self-studying with monthly revision workshops. The self-studying modules were incentivized CAL-assignments based on learning videos and quizzes developed by Khan Academy and administered via the offline application Kolibri. To measure teachers' content knowledge, we conducted assessments based on the local primary school math curriculum before, shortly after, and one year after the program. An initially planned student assessment one year after the program's conclusion could not be carried out due to the Covid-19 pandemic.

We find that immediately after the intervention, program teachers outperformed their peers from the control group by 0.29σ ($p<0.01$) or 5.52 percentage points ($p<0.01$), but this effect diminished by 72 percent one year later. The data further reveals considerable heterogeneity in treatment effects. In the short term, the program was particularly successful in raising test scores among teachers under 40 (0.53σ, $p<0.01$), and regarding more advanced concepts from grades five to six (0.31 to 0.35σ, $p<0.01$), but even these effects became insignificant after one year.

Investments in teacher competencies have the potential to be highly cost-effective. If teachers retain their acquired skills, further student cohorts will benefit after the intervention period. This stands in contrast to student-centered interventions such as remedial CAL classes that often require continued investments. A unique feature of this study is that the results from the CAL-based teacher training can be directly compared to findings from student-centered CAL lessons implemented in the same context and by the same NGO. Based on the parameters obtained through our experiment, we simulate the long-term cost effectiveness of our teacher intervention and compare it to the cost-effectiveness of remedial CAL lessons we experimentally evaluated with third to sixth graders (see Büchel et al., 2021). Our benchmark findings indicate that an annual retention rate of at least 55 percent among treated teachers is required so that the CAL training with teachers would be more effective than CAL lessons with students. In our experiment, we observe a retention rate of 28 percent, suggesting that the long-term effectiveness

of the teacher program is lower than that of the student intervention.

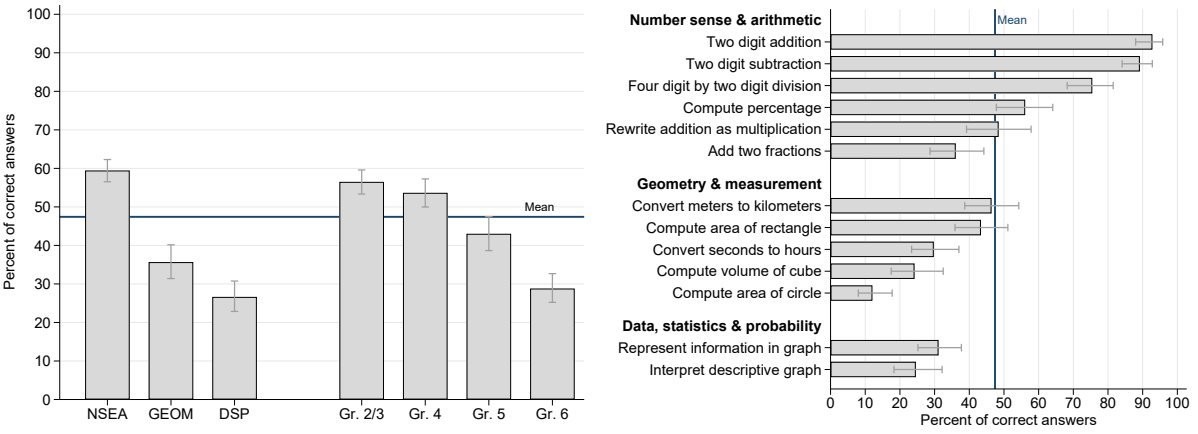
The high depreciation rate of effects at the teacher level is in line with the sparse evidence on this topic. Bando and Li (2014) find substantial gains in competencies of Mexican teachers from a intensive training on English skills and instructional methods, but the gap between the treatment group and the control group faded after 12 months. Similarly, Cilliers et al. (2019, 2020) report substantial short term gains for two pedagogy-centered teacher training programs, but only for one of the programs these effects were found to persist. Our study complements these findings by showing that steep depreciation rates in newly acquired skills are also a key challenge in purely content-related teacher training programs focusing on primary school mathematics. Considering the lack of evidence on the sustainability of professional development programs and the relevance of the topic for educational policy, this likely remains an important avenue for future research.

2 Assessing Teacher Content Knowledge in El Salvador

Despite impressive improvements in the accessibility of primary education, the quality of schooling often remains alarmingly low in developing countries. According to statistics by the World Bank (2018), less than 40 percent of students in a typical lower-middle income country pass minimum thresholds in mathematics by the end of primary school, and this rate drops to 14 percent in low income countries.

While many features of the schooling system affect the learning achievements of students, teachers are widely considered the most important input to the educational production function (Baumert and Kunter, 2013; Hanushek, 2011), and their salaries account for the bulk of education spending (Bold et al., 2017b). Barber and Mourshed (2007) conclude from their study of high-performing school systems that "the quality of an education system cannot exceed the quality of its teachers". Hence, it is critical to understand how well teachers are prepared for the challenging task that awaits them in the classroom. One essential pre-requisite for effective teaching is a sound mastery of the concepts to be taught. However, recent evidence from African countries and India points to alarmingly low levels of teacher content knowledge. Most notably, Bold et al. (2017a) report results on teacher skills based on a large-scale assessment across seven countries in Sub-Saharan Africa. According to their definition, a teacher possesses minimum subject knowledge in mathematics if she or he is able to mark at least 80 percent of items on a mock test for forth graders correctly. On average, only two thirds of the teachers met this low requirement, with estimates ranging from 93 percent in Kenya to 49 percent in Togo. They find that deficiencies in teachers' content knowledge account for 30 percent of the shortfalls in student learning relative to the curriculum, and about 20 percent of the cross-country difference in student performance in their sample. Similar results are reported for the Indian province Bihar, where only 34 percent of the teachers were able to solve a perimeter problem corresponding to grade five (Sinha et al., 2016).

Our research adds to the still sparse evidence on teachers' content knowledge by conducting a representative assessment in the department of Morazán in El Salvador, a lower middle-income country in Central America. According to recent World Bank data, both the access to and the quality of primary schooling in El Salvador is below the average for lower middle-income



(a) Share of correct answers by domain and grade (b) Share of correct answers for selected items

Figure 2: Content knowledge of math teachers in Morazán, El Salvador
Note: N= 224; survey design taken into account; capped spikes indicate 95% confidence intervals. Results in subfigure 2a are divided into the domains Number Sense and Elementary Arithmetic (NSEA), Geometry and Measurement (GEOM) and Data, Statistics and Probability (DSP).

countries.[2] Morazán is located in southeastern El Salvador and is one of the poorest regions of the country. In national assessments, its secondary students perform roughly at the country average.[3]

For the first part of our study, we randomly sampled 231 math teachers from 107 public primary schools and asked them to participate in an exam-type assessment. Overall, 224 teachers (97%) complied with our invitation and took part in a 90-minute paper-and-pencil test comprising 50 items from the Salvadorian primary school math curriculum. The weighting of questions across the three domains *Number Sense and Elementary Arithmetic* (∼65%), *Geometry and Measurement* (∼30%), and *Data, Statistics and Probability* (∼5%) was closely aligned with the national curriculum. The test covered concepts taught in grade 2 (6 items), grade 3 (13 items), grade 4 (10 items), grade 5 (11 items), and grade 6 (10 items). To make sure that the items were suitable for the Salvadorian context, the assessment was reviewed by local teaching experts and the local education ministry.[4]

Figure 2 presents the results by subject domain and item difficulty (subfigure 2a) and for selected example items (subfigure 2b). The average teacher is able to answer 47 percent of grade two to six questions correctly, and performance is poor across all tested subject domains. Learning shortfalls are most apparent in *Data, Statistics and Probability* (27% correct answers) and *Geometry and Measurement* (36% correct answers), and least pronounced regarding *Number Sense and Elementary Arithmetic* (59% correct answers). Many teachers not only struggle with the more advanced items pertaining to grade six (29% correct answers), but even with items covering the basic materials from grades two and three (57% correct answers). While most

[2]Measures for learning outcomes based on various international assessments at the primary school level were recently harmonized by Angrist et al. (2021). These harmonized learning outcomes are provided online by the World Bank, as are net primary school enrollment statistics: `https://data.worldbank.org/indicator`.

[3]In 2019, Morazán ranked seventh among the 14 Salvadoran departments in the "PAES" examination, a standardized test administered to all secondary school students throughout the country (MINED, 2019).

[4]More information on the sampling of participants and the assessment design is provided in the Appendix sections B.1 and B.3. We discuss additional results from the mathematics assessment along the opinion of teachers collected via a survey in Brunetti et al. (2020).

teachers can handle basic operations such as additions or subtractions (about 90%), only 56% are able to solve a simple operation involving percentages, less than half (46%) can convert meters to kilometers or compute the area of a rectangle (43%), about a third can add two fractions (36%), and only one in four (25%) can retrieve information from a descriptive chart. Applying the minimum proficiency threshold advocated by Bold et al. (2017a), our assessment suggests that only 14 percent of teachers possess sufficient content knowledge to effectively teach math at the primary school level.

These results are particularly striking given that 97 percent of the teachers in our sample possess a university degree, meaning that they have either completed a teaching degree (2 to 3 years, 70% of teachers) or a bachelor's degree (5 to 6 years, 27% of teachers). Hence, despite 13 to 17 years of formal education, most primary schools teachers are confronted with the daunting task of teaching what they don't know. If quality education for all is to be achieved, it is thus critical to find effective ways to sustainably improve teacher skills.

3 Improving Teacher Content Knowledge through Computer-Assisted Learning

3.1 Intervention

For the second part of this study, we cooperated with the Swiss-Salvadoran NGO Consciente to implement an in-service teacher training program between April and August 2019. The intervention targeted 87 primary school math teachers and consisted of two elements: *(i)* computer-assisted self-studying at home, and *(ii)* monthly revision workshops.

Self-Studying. Drawing on the extensive materials of the learning software Khan Academy, 16 study modules covering selected contents of the Salvadoran primary school math curriculum were designed. In accordance with the official curriculum, the main focus of the training program was on *Number Sense and Elementary Arithmetic*, but concepts pertaining to *Geometry and Measurement* and *Data, Statistics and Probability* were covered as well. In an initial meeting, participants received a laptop equipped with the learning software, which allows offline access to the selected learning videos and exercises from Khan Academy.[5] Teachers had to complete one module per week, corresponding to a workload of four to eight hours, and then took a short assessment administered by the software. Since module completion had to be accomplished outside working hours, teachers received monetary compensation for it. Payments were conditional on the completion of the assigned exercises and videos (weight: 0.85) and on quiz performance at the end of each module (weight: 0.15). For the first module, teachers could earn up to 18.00 USD. In terms of Salvadoran wage levels, this roughly corresponds to a regular teacher salary for half a workday. With each subsequent module, maximum compensation increased by 0.50 USD yielding 25.50 USD for the final assignment. Throughout the interven-

[5] Khan Academy is free of charge and features math content in more than 30 languages. The full version is available in 16 languages including Spanish, and a subset of content is available in about 20 languages. Like in many developing countries, poor internet coverage is a challenge in El Salvador. We therefore deployed an open-source platform, Kolibri, designed to make offline learning with content from Khan Academy and other CAL-sources possible.

tion, the software monitored teachers' progress and participants received regular reminders and individual support in case of technical problems.

Monthly Workshops. At the monthly workshops, participants submitted the work they accomplished on the previous four self-studying modules. While teachers took part in a tutoring session, their learning progress in the self-studying modules was evaluated to determine the compensation they were to receive. During the workshops, expert teachers recapitulated key concepts and addressed teachers' questions. Meetings were scheduled for half a day and, as they took place during work hours, teachers were only compensated for travel expenses.

3.2 Experimental Design

To evaluate the impact of the program on teachers' math performance, we set up a randomized controlled trial, where applicants to the teacher training program were randomly assigned to either the treatment or the control group. Before, shortly after, and one year after the intervention teachers were administered a comprehensive math assessment. A comparison between the two groups allows us to track the causal effect of the program on teacher content knowledge over time.

Sampling and Randomization. To recruit the study participants, our partner NGO sent out invitations to grade three to grade six math teachers in 253 schools throughout Morazán. In total, 313 teachers from 175 different schools applied for the program. After a baseline assessment, the worst-performing applicant in each school was selected for study participation, yielding a final sample of 175 teachers from 175 different schools.[6] Finally, the 175 pre-selected teachers were randomly assigned to either the control group (88 teachers) or the treatment group (87 teachers). To enhance the efficiency of the estimates, randomization was stratified by baseline score and gender.

Data and Measurement. The teacher assessments comprised 50 items from various sources and were designed to mirror the Salvadoran math curriculum for grades two to six (see appendix B.4 for more information). The assessments were administered during regional teacher meetings and had to be completed in 90 minutes. We further collected data on teacher characteristics through a brief survey we administered directly after the baseline assessment. Our teacher data is complemented by administrative data on school characteristics provided by the education ministry as well as monitoring data on module completion and workshop attendance collected during the intervention.

Baseline Characteristics. Table A.1 in the appendix shows that baseline characteristics are well-balanced across the two experimental groups. In both the treatment and the control group, the average teacher scored 43 percent correct answers and is thus slightly below the the regional average of 47 percent (see section 2). Moreover, the average teacher in our sample is 44 years old,

[6]Note that this part of the sampling procedure was not communicated to applicants to avoid misaligned incentives during the assessments.

64 percent of the study participants are female and all of them except one completed tertiary education.

Compliance and Attrition. Good completion rates for modules (74%) and high attendance rates at workshops (85%) show that teachers complied well with the experimental protocol (see Figure A.1 in the appendix). While all 175 teachers participated in the baseline assessment, 164 teachers took the endline assessment shortly after the intervention (6% attrition), and 136 teachers participated in the follow-up assessment one year later (22% attrition). Table A.2 compares attrition rates across experimental groups and provides no indication that participation in assessments correlated significantly with the treatment status.

3.3 Results

We estimate the intent-to-treat (ITT) treatment effect of being randomly assigned to the treatment group at endline (i.e., *EL*=one month after the intervention) or follow-up (i.e., *FU*=one year after the intervention) based on

$$Y_{jk}^{wave} = \alpha + \beta\, Treat_{jk} + \delta Y_{jk}^{BL} + X_{jk}^{'}\gamma + S_{jk}^{'}\rho + \phi_k + \epsilon_{jk} \quad \text{for } wave \in [\text{EL, FU}], \tag{1}$$

where Y_{jk}^{wave} represents the endline (or follow-up) math score of teacher j in stratum k and is either measured as the percentage share of correct answers or the standardized share of correct answers such that the control group's mean in a given wave is zero ($\mu_{control}^{wave} = 0$) and the standard deviation is one (i.e., $\sigma_{control}^{wave} = 1$). The main variable of interest is the binary indicator $Treat_{jk}$ that equals one if teacher j belongs to the treatment group. Y_{jk}^{BL} denotes the baseline test score and X_{jk} are additional pre-determined teacher attributes including age, gender, highest educational degree, years since graduation, math specialization and commuting time to school. S_{jk} captures covariates at the school level including an equipment and an infrastructure index, travel time to the department's capital as well as binary indicators for the availability of a computer lab, gang activities on school grounds and location in a rural area. Finally, ϕ_k denotes stratum fixed effects and ϵ_{jk} is the error term. In the following, we report results based on equation (1) as well as a sparse specification excluding the pre-determined teacher attributes X_{jk} and school level characteristics S_{jk}.

Immediate Program Effect. Table 1 displays the benchmark estimates for the effect of the program on teachers' content knowledge.[7] In columns (1) to (4), we estimate the short-term program effects measured one month after the program ended. Columns (1) and (2) show that the evaluated teacher training program raised the share of correct answers by 5.38 to 5.52 percentage points (p-value<0.01). This translates to an impact of 0.28σ to 0.29σ (p-value<0.01) when the program effect is estimated based on standardized scores as in columns (3) and (4).

An authoritative assessment of our results against previous findings is difficult because only few studies quantify the impact of teacher training programs on teacher content knowledge:

[7] In the appendix section A.4, we present density plots for the participants' share of correct answers at the baseline, endline, and follow-up assessments disaggregated by treatment status.

Table 1: ITT-estimates for the program effects on teachers' math scores

	Immediate effect				Effect after one year			
Dependent variable:	Percent correct		Standardized		Percent correct		Standardized	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	5.38***	5.52***	0.28***	0.29***	0.61	1.48	0.03	0.08
	(1.46)	(1.49)	(0.08)	(0.08)	(1.78)	(1.77)	(0.10)	(0.10)
Baseline score	0.90***	0.85***	0.92***	0.86***	0.77***	0.64***	0.82***	0.68***
	(0.09)	(0.10)	(0.09)	(0.10)	(0.12)	(0.13)	(0.13)	(0.14)
Adjusted R^2	0.80	0.81	0.80	0.81	0.69	0.71	0.69	0.71
Observations	164	164	164	164	136	136	136	136
Teacher controls[a]	No	Yes	No	Yes	No	Yes	No	Yes
School controls[b]	No	Yes	No	Yes	No	Yes	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *immediate effect* is estimated based on the endline data collected in September 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in September 2020. In *columns (3), (4), (7), and (8)*, the share of correct answers is standardized to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. *a: Teacher level controls* include age, educational degree, years since graduation, commuting time to school as well as binary indicators for gender and math specialization. *b: School level controls* are an infrastructure index, an equipment index, travel time to the department's capital as well as binary indicators for the availability of a computer lab, exposure to gang activities, and location in a rural area. Huber-White robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Bando and Li (2014) report standardized treatment effects of 0.35σ on teachers' English proficiency from a six-month professional development program in Mexico, whereas Zhang et al. (2013) find no significant impact on teachers' English skills from an intensive three-week program in Chinese migrant schools. Taking experimental impact evaluations on children's learning outcomes as a benchmark, the program's immediate impact of 0.29σ on teachers' content knowledge is sizeable. Even for well-proven types of educational interventions, such as remedial education or computer-assisted learning, systematic reviews report average effect sizes on children's math scores below 0.2σ (e.g., Snilstveit et al., 2015; McEwan, 2015).

Persistency of the Program Effect. How persistent is the program effect in the long run? Columns (5) to (8) of Table 1 indicate that less than one third of the impact remains after one year. The effect estimates based on the follow-up assessment vary between 0.6 percentage points (column 5, no controls) and 1.5 percentage points (column 6, with controls) or 0.03σ (column 7, no controls) and 0.08σ (column 8, with controls) and are imprecisely estimated with p-values between 0.40 and 0.73. Hence, the reported immediate gains in teachers' content knowledge were rather elusive, as the short-term effect depreciated by more than two-thirds after one year.

How do these findings compare to other studies? The evidence base on the long-term sustainability of teacher training programs is still surprisingly scarce, but so far it confirms that achieving persistent effects through teacher training programs is challenging. In line with our results, Bando and Li (2014) find that the gap in English proficiency between participants and the control group documented immediately after the intervention disappeared when they reassessed the Mexican teachers twelve months later. Research by Cilliers et al. (2019, 2020) examines

the sustainability of two teacher development programs focusing on teaching techniques instead of content knowledge. Their findings suggest that professional development programs are able to produce sustainable improvements among participants, but that persistent program effects cannot be taken for granted, even when a sizable short-term impact has been achieved.[8] Similar to Cilliers et al. (2019, 2020) and Bando and Li (2014), the results presented in Table 1 underscore that short-term gains do not necessarily translate to a sustained impact that persists in the long-run.

Effect Heterogeneity and Robustness. To gain a more nuanced understanding of the program's impact on teachers, we first explore several dimensions of effect heterogeneity and then asses whether the follow-up results may be driven by selective attrition.

Table A.3 in the appendix shows that the immediate effect as well as the persistency of the impact did not vary by item domain: The teacher training program was equally effective in producing short-term gains in *Number Sense & Elementary Arithmetic* (0.25σ, $p<0.01$) and in *Geometry, Measurement, Data & Statistics* (0.30σ, $p<0.01$), and the effects across both domains largely disappear after one year. While we find no heterogeneity along domain, Table A.4 shows that the program was about twice as effective at improving the participants' proficiency in concepts from grade levels five and six (0.31–0.36σ, $p<0.01$) compared to concepts from grade levels three and four (both 0.19σ, $p=0.06$–0.13). Yet, the ITT-estimates become insignificant across items of all grade levels at the follow-up assessment one year after the conclusion of the program.

Testing for effect heterogeneity along teacher characteristics in Table A.5 and Figure A.3 shows that older teachers were significantly less perceptive than their younger colleagues. For instance, participants older than 50 gained on average 0.04σ ($p=0.81$) at endline, while their youngest colleagues (≤ 40) experienced average gains of 0.53σ ($p<0.01$); the gap between these two age groups is significant at the 5% level (p-value=0.02). But even for teachers younger than 40, we do not obtain a significant program impact after one year (0.15σ, $p=0.42$). We also find that participants with the lowest baseline score gained the least, but the effect differences along baseline ability are not statistically significant.

While the data reveal effect heterogeneity across several dimensions, all of the sub-analyses replicate the substantial deterioration in program effects after 12 months. Since only 136 teachers participated in the follow-up assessment (compared to 175 teachers at baseline and 164 teachers at endline), one may be concerned about bias induced by selective attrition. To understand the relevance of potential selection effects at the follow-up assessment, Table A.6 restricts the analysis to the sub-sample of 131 teachers that participated in all three assessments. This leaves the point estimates unaltered compared to the benchmark analysis. For instance, the full

[8] One intervention arm studied by Cilliers et al. (2019, 2020) was delivered in the form of a four days training workshop designed to demonstrate how participants can teach a language and literacy curriculum effectively (training-based approach), while the second intervention arm was built around monthly visits from coaches who provided feedback on the participants' pedagogical techniques (coaching approach). For the *training-based* intervention, the authors report a substantial fade-out for the program's effect on teacher behavior: Depending on the measured outcome, the program's effects declined by 50% to 90% over one year and became statistically insignificant. The *coaching-based* intervention was more successful in producing sustainable impacts on teacher behavior, as between 66% and 100% of the immediate program effects carried over to the follow-up survey after one year.

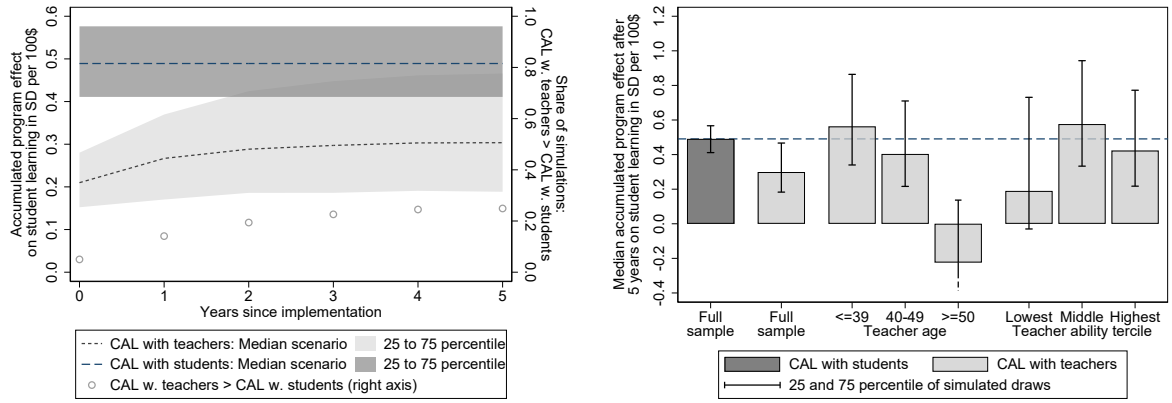
specification for standardized scores at endline yields an effect of 0.29σ ($p<0.01$) in column (4) of Table 1 compared to 0.28σ ($p<0.01$) based on the restricted sample in column (4) of Table A.6. Similarly, the changes in impact estimates for standardized scores after one year are negligible when we use the restricted sample (0.08σ, $p=0.40$) instead of the benchmark specification (0.09σ, $p=0.38$). Finally, Table A.7 re-estimates the benchmark specification weighting observations by their inverse probability of selection into the endline or follow-up assessment. To be precise, we use entropy balancing (see Hainmueller, 2012; Jann, 2021) to reweight both the treatment group and the control group in a way such that the distribution of covariates in both groups (and during all assessments) is identical to the pooled distribution in the overall sample of 175 teachers. Again we do not find any indication that the deterioration in program effects after one year are driven by selective attrition along observable attributes.

3.4 Discussion: Learning Gains among Students and the Cost-Effectiveness of the Program

Two questions that naturally arise are *(a)* to what extent increased content knowledge of teachers transmits to their students, and *(b)* whether it is more cost-effective to organize CAL sessions for students or their inadequately prepared teachers. Originally, the field experiment was designed to address both questions *directly*, but school closures as a response to the COVID-19 pandemic disrupted in-class transmission from teachers to students for several months and infection risks inhibited large-scale assessments with children.

Teacher Content Knowledge & Learning Gains among Students. Due to COVID-19 related constraints, we infer the program's impact on students using observational Salvadoran data and quasi-experimental estimates from related research. Our teacher data can be combined with information on student learning from a field experiment conducted in the same context (see Büchel et al., 2021), allowing us to estimate the impact of teachers' content knowledge on learning gains of their students over the course of a school year. Our results suggest that a 1σ better teacher knowledge is associated with a 0.09σ to 0.12σ gain in student learning (for details, see appendix section A.9). These observational estimates for El Salvador are closely aligned with quasi-experimental evidence for Peru (Metzler and Woessmann, 2012) and Pakistan (Bau and Das, 2020), suggesting a transmission parameter of 0.09 between teacher content knowledge and student learning in mathematics (see Table A.9 for a summary of evidence). Applying this transmission parameter, the immediate program effect on teachers' content knowledge of 0.29σ translates to a very small gain in average student learning of 0.026σ (i.e., $0.09 \times 0.29\sigma$), which is equivalent to 0.08 additional years of schooling (for details, see appendix section A.9).

Cost-Effectiveness of the Program. A fundamental advantage of teacher training programs are potential long-term cascade effects: CAL interventions targeting students require the *continuous* maintenance of large computer labs, whereas improving *one* teacher's content knowledge enhances the learning experience of *many* children *every* year. This brings about two favorable implications: First, the program costs per (indirectly) targeted child during a teacher training are considerably lower than the program costs per child for additional CAL lessons. Second, the



(a) Simulations based on full sample estimates (b) Impact after 5 years for various target groups

Figure 3: Simulated cost-effectiveness of CAL with teachers and CAL with students
Note: Figure 3a is based on 1000 random draws using the following parameters: Distribution of immediate program effect based on Table 1 $\sim \mathcal{N}(0.29,\ 0.08^2)$; distribution of program effect after one year based on Table 1 $\sim \mathcal{N}(0.08,\ 0.10^2)$; covariance between immediate and follow-up effect=0.004; distribution of effect transmission from teachers to students based on Tables A.8 and A.9 $\sim \mathcal{N}(0.09,\ 0.03^2)$; program costs of CAL directed at teachers=12$ per student; distribution of annual effect of CAL program directed at students based on Büchel et al. (2021) $\sim \mathcal{N}(0.21,\ 0.05^2)$; program costs of CAL directed at students based on Büchel et al. (2021)=43$ per student. *Figure 3b* depicts the accumulated effect on students after five years for scenarios targeting different teacher groups that were analyzed in Section A.6: the plotted bars correspond to the median value after 5 years, while the capped spikes reproduce the 25th and 75th percentiles represented as shaded areas in Figure 3a.

costs of additional CAL lessons to children accrue periodically, while a one time investment in teacher skills produces recurrent gains – although these likely fade out as the treatment effect on teachers' content knowledge depreciates.

With these considerations in mind, we calculate the cost-effectiveness of the CAL-based teacher training combining four elements: *(i)* the immediate impact of the CAL training on teacher content knowledge, namely estimates from column (4) of Table 1; *(ii)* the annual depreciation in the program effect on teacher content knowledge combining the long-term effect estimates in column (8) of Table 1 with the immediate impact estimates; *(iii)* one-time implementation costs per (indirectly targeted) student calculated to 12 USD using the guidelines by Dhaliwal et al. (2014); *(iv)* the transmission of teacher content knowledge to students' learning gains, as discussed above.

One particularly valuable feature of this study is that its results can be compared to a companion paper evaluating CAL lessons offered to pupils of grades three to six (see Büchel et al., 2021). Importantly, the two field experiments were conducted in the same environment (i.e., primary schools in the Salvadorian department Morazán), using the identical CAL software for teaching basic mathematics (i.e., Khan Academy content via an offline application), and both interventions were implemented together with the same partner organization (i.e., the Swiss-Salvadoran NGO Consciente).

Figure 3a depicts cost-effectiveness estimates from Büchel et al. (2021) and cost-effectiveness estimates for the CAL-based teacher training based on 1000 random draws. We model uncertainty by drawing each cost-effectiveness parameter (except the implementation costs per student) from a normal distribution with a mean equal to the parameter's point estimate and a variance equal to the point estimate's squared standard error. The left-hand axis shows the

accumulated program effect on student learning (measured in σ) per 100 USD, whereas the right-hand axis depicts the share of simulations yielding larger accumulated effects for CAL-based teacher trainings than for CAL-based lessons directed at students. The results suggest that the evaluated CAL lessons for students were likely more cost-effective than the evaluated CAL-based teacher trainings. In the median scenario, a 100 USD investment in CAL lessons for students increases learning gains by 0.49σ compared to 0.31σ for the same 100 USD investment in CAL-based teacher trainings. These impact estimates correspond to 1.5 school year equivalents for CAL directed at students, and 1 school year equivalent when CAL training is provided to teachers (for details, see appendix section A.9).

The shaded areas in Figure 3a represent draws between the 25th and 75th percentile, and indicate that the variance in the cost-effectiveness estimates for CAL-based teacher trainings is substantial. Yet, even when taking this large variance into account, CAL-based teacher trainings outperform CAL lessons for students in only 25 percent of the simulated scenarios, as depicted by the gray hollow circles. Most of the uncertainty in the cost-effectiveness simulations for the teacher trainings arises from the imprecise estimates on the persistency of program effects. If we remove this uncertainty by feeding the simulation with constant follow-up estimates (i.e., mean=0.08 and sd=0), the share of draws where CAL-based teacher trainings outperform additional CAL lessons for students decreases from 25 to 14 percent. To make the CAL training with teachers at least as cost-effective as CAL lessons with students, our simulations suggest that a retention rate of 55 percent among treated teachers would be required, which is twice as high as the retention rate observed in the experiment.

While these results are not in favor of software-based professional development programs, systematically targeting the most perceptive teachers would likely improve the cost-effectiveness of the policy. Figure 3b presents simulation results for the accumulated program effect on student learning after 5 years under different targeting regimes. The plotted bars correspond to the median scenario after 5 years, and the capped spikes reproduce the 25th and 75th percentiles represented as shaded areas in Figure 3a. The results in Figure 3b highlight that targeting young and middle-aged teachers would likely lift the teacher training's cost-effectiveness close to or even beyond the cost-effectiveness of additional CAL lessons for students. The same conclusion applies to targeting teachers in the second and third ability terciles. Having said that, a better understanding on how to achieve more persistent impacts in teacher trainings is arguably the key to unlocking the policy's full potential and increasing cost-effectiveness manifold.

4 Conclusion

Well qualified teachers are an essential requirement to achieve *quality education for all*, as envisioned by the *2030 Agenda for Sustainable Development*. Drawing on data from a representative math assessment, this study documents that primary school math teachers in northeastern El Salvador only master 47 percent of the curriculum they teach. This number is based on a direct assessment of teacher skills and is considerably lower than previous estimates for other developing countries relying on an indirect assessment through the grading of mock student tests.

Our field experiment shows that targeted teacher training using CAL software can produce

substantial short-term gains in teachers' content knowledge. After a five-month teacher training program, we observe an average intention-to-treat effect of 0.29σ, with estimates ranging from effectively zero for teachers over the age of 50 to 0.52σ for teachers under 40. However, achieving sustained improvements in teacher skills proved to be more challenging. Learning gains at the teacher level depreciate by 72 percent to a mere 0.08σ one year after the treatment.

The unique setting of our experiment allowed us to compare the cost-effectiveness of CAL for teachers with that of an analogous CAL experiment directly targeting students. Teacher-centered initiatives are generally seen as a highly sustainable educational investment because they potentially benefit all future student cohorts a teacher instructs. Our simulations suggest that this assumption only holds if learning gains at the teacher level can be largely maintained over time. Based on the empirical parameters of the two experiments, we estimate that the retention rate of the effect on teacher knowledge should be at least 55 percent to guarantee that CAL for teachers is more cost-effective than CAL for students. With the actual retention rate of 28 percent we observed in our teacher experiment, the student-centered approach can be considered superior.

Our findings illustrate the importance of going beyond short term gains when evaluating the effectiveness of policies and interventions. While some programs can only be expected to have an impact on the cohort that was directly treated, others may induce sustained changes that can substantially increase the overall cost-effectiveness. Future research should appreciate this and help identify effective ways of ensuring the persistency of the achieved gains.

References

Angrist, Noam, Simeon Djankov, Pinelopi Goldberg, and Harry Patrinos. 2021. Measuring human capital using global learning data. *Nature* 592:403–408.

Bando, Rosangela and Xia Li. 2014. The effect of in-service teacher training on student learning of English as a second language. IDB Working Paper Series No. 529.

Barber, Michael and Mona Mourshed. 2007. How the world's best-performing schools systems come out on top. Mc Kinsey Report.

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

Baumert, Jürgen and Mareike Kunter. 2013. The COACTIVE Model of Teachers' Professional Competence. In *Cognitive activation in the mathematics classroom and professional competence of teachers*, ed. Mareike Kunter et al. New York: Springer, 25–48.

Bietenbeck, Jan, Marc Piopiunik, and Simon Wiederhold. 2018. Africa's skill tragedy: Does teachers' lack of knowledge lead to low student performance? *Journal of Human Resources* 53 (33):553–578.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian Stacy, Jakob Svensson, and Waly Wane. 2017b. What do teachers know and do? Does it matter? Evidence from primary schools in Africa. World Bank Policy Research Working Paper No. 7956.

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

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

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

Büchel, Konstantin, Martina Jakob, Kühnhanss Christoph, Daniel Steffen, and Aymo Brunetti. 2021. The relative effectiveness of teachers and learning software. evidence from a field experiment in El Salvador. *Journal of Labor Economics,* forthcoming, `https://doi.org/10.1086/717727`.

Cilliers, Jacobus, Brahm Fleisch, Janeli Kotze, Mpumi Mohohlwane, and Stephen Taylor. 2019. The challenge of sustaining effective teaching: Spillovers, fade-out, and the cost-effectiveness of teacher development programs. *Unpublished manuscript.*

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor. 2020. How to improve teaching practice? an experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources* 66:203–213.

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

Escueta, Maya, Andre Nickow, Philip Oreopoulos, and Vincent Quant. 2020. Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature* 58 (4):897–996.

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

Hainmueller, Jens. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20 (1):25–46.

Hanushek, Eric. 2011. The Economic Value of Higher Teacher Quality. *Economics of Education Review* 30 (3):466–479.

Jackson, Kirabo, Jonah Rockoff, and Douglas Staiger. 2014. Teacher effects and teacher-related policies. *Annual Review of Economics* 6 (1):801–25.

Jann, Ben. 2021. Entropy balancing as an estimation command. University of Bern Social Sciences Working Paper No. 39. Online available, URL: `https://boris.unibe.ch/157883/`.

Kerwin, Jason and Rebecca Thornton. 2021. Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures. *Review of Economics and Statistics* 103 (2):251–264.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. The challenge of education and learning in the developing world. *Science* 340 (6130):297–300.

McEwan, Patrick. 2015. Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research* 85 (3):353–394.

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

MINED, Ministerio de la Educación Ciencia y Tecnología de El Salvador. 2019. Informe de resultados: Paes 2019. Online avaiable, URL: `https://www.mined.gob.sv`.

Rodriguez-Segura, Daniel. 2021. EdTech in developing countries: A review of the evidence. *The World Bank Research Observer* `https://doi.org/10.1093/wbro/lkab011`.

San Antonio, Diosdado, Nelson Morales, and Leo Moral. 2011. Module-based professional development for teachers: a cost-effective Philippine experiment. *Teacher Development* 15 (2):157–169.

Sinha, Shabnam, Rukmini Banerji, and Wilima Wadhwa. 2016. *Teacher performance in Bihar, India: Implications for education.* Washington D.C.: The World Bank.

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.

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

Zhang, Linxiu, Fang Lai, Xiaopeng Pang, Hongmei Yi, and Scott Rozelle. 2013. The impact of teacher training on teacher and student outcomes: Evidence from a randomised experiment in Beijing migrant schools. *Journal of Development Effectiveness* 5 (3):339–358.

A Appendix: Analysis

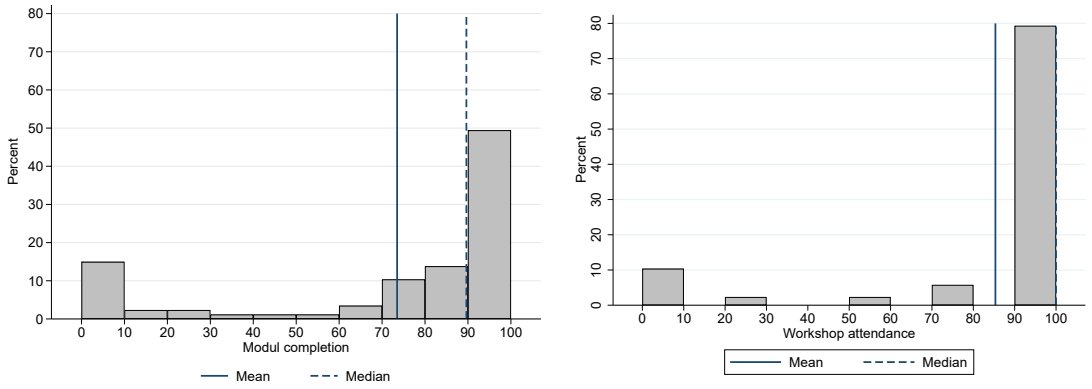
A.1 Characteristics at Baseline

Table A.1: Baseline characteristics by treatment status

	Treatment group	Control group	p-value
Panel A: Baeline math scores (N=175)	(1)	(2)	(3)
%-Share correct answers	43.26	43.27	1.00
	(2.94)	(2.07)	
Standardized math score	0.00	0.00	1.00
	(0.15)	(0.11)	
Baseline test group: March 2019[a]	0.32	0.36	0.56
	(0.07)	(0.05)	
Panel B: Sociodemographics (N=175)			
Age	44.36	43.78	0.64
	(1.21)	(0.85)	
Female	0.64	0.64	0.92
	(0.07)	(0.05)	
Highest degree[b]	2.23	2.25	0.76
	(0.06)	(0.05)	
Year since highest degree	19.77	18.82	0.44
	(1.22)	(0.86)	
Math specialization[c]	0.08	0.06	0.54
	(0.04)	(0.03)	
Travel time to school (min.)	58.80	72.28	0.17
	(9.86)	(6.95)	
Panel C: School level information (N=175)			
Computer access students	0.46	0.38	0.26
	(0.07)	(0.05)	
Equipment index[d]	0.27	0.26	0.63
	(0.03)	(0.02)	
Infrastructure index[d]	0.27	0.27	0.89
	(0.02)	(0.02)	
Gang activities on school grounds	0.11	0.09	0.60
	(0.05)	(0.03)	
Rural area	0.86	0.85	0.85
	(0.05)	(0.04)	
Travel time to department capital (min.)	47.70	50.22	0.56
	(4.26)	(3.00)	

Notes: This table presents the mean and standard error of the mean (in parentheses) for baseline characteristics by treatment status. Column 3 shows the p-value (based on two-sided t-tests) from testing whether the mean is equal across control and treatment group. *a:* A dummy variable indicating whether the teacher took the baseline in September 2018 (0) or March 2019 (1). *b*: Four categories: 1=bachillerato (high school), 2=profesorado (2–3 years of tertiary education), 3=licenciatura (5–6 years of tertiary education, equiv. to a bachelor's degree) and 4=maestria (equiv. to a master's degree); since bachillerato (1→2) and masteria (4→3) occur only once in the estimation sample, we reassign them to the adjacent category. *c*: Respondent teaches math only (1) or various subjects (0). *d*: For each school a list covering twelve technical equipments and eleven facilities is available. The equipment and infrastructure indices refer to the share of items or facilities on this list that a school possesses.

A.2 Program Compliance



(a) Completion rate of modules (b) Attendance rate of meetings

Figure A.1: Compliance of teachers with the program
Note: The overall compliance rate is the weighted average of the module completion rate in Figure A.1a and the attendance rate in Figure A.1b with an average of 75 percent and a median of 91 percent.

A.3 Attrition

Eleven teachers (6.3%) did not take part in the endline assessment, and 39 teachers (22.2%) missed the follow-up assessment conducted during the first year of the COVID-19 pandemic. Table A.2 reports estimates from linear probability models (LPM) that test whether the attrition status of participants is correlated with the treatment assignment. The estimates in columns (1) to (3) include different set of control variables and unambiguously suggest that attrition at the endline assessment is uncorrelated with treatment status (p= 0.56−0.78). Similarly, columns (4) to (6) yield an insignificant correlation between the treatment status and attendance at the follow-up assessment (p= 0.56 − 0.60). The same conclusions hold if the correlation between treatment status and attrition is estimated with a Logit model (results not shown).

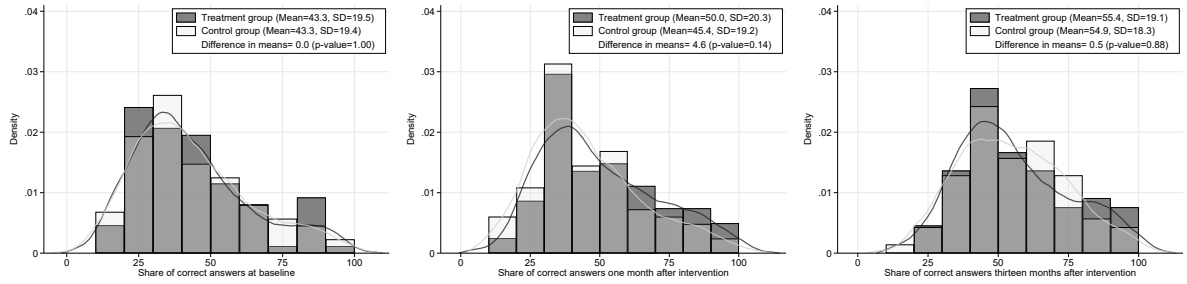
Table A.2: Linear probability model for attrition by treatment status

Attrition status at:	Endline assessment			Follow-up assessment		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.012	0.011	0.023	0.037	0.033	0.020
	(0.037)	(0.039)	(0.039)	(0.063)	(0.064)	(0.064)
Baseline score		−0.003*	−0.004		0.000	−0.001
		(0.002)	(0.002)		(0.005)	(0.006)
Observations	175	175	175	175	175	175
Teacher controls	No	No	Yes	No	No	Yes
School controls	No	No	Yes	No	No	Yes
Stratum FE	No	Yes	Yes	No	Yes	Yes

Notes: Huber-White robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

A.4 Distribution of Correct Answers by Wave and Treatment Status

Figure A.2 presents density plots for the participants' share of correct answers at the baseline, endline, and follow-up assessments disaggregated by treatment status. Before the implementation of the program in spring 2019, the distributions of correct answers given by the treatment group and by the control group closely coincide (difference in means=0.0, p-value=1.00). At the endline assessment, about one month after the professional development program ended, we observe an increase in the share of correct answers in the treatment group compared to the control group. The difference in means at endline is 4.6 percentage points with a p-value of 0.14. One year later, at the follow-up assessment, the two distributions again largely overlap with a difference in means of 0.5 percentage points (p-value=0.88).



(a) Baseline (March 2019) (b) Endline (Sept. 2019) (c) Follow-up (Sept. 2020)

Figure A.2: Share of correct answers by treatment assignment

A.5 Program Effects by Subtopic

Table A.3: ITT-Estimates on the effects on teacher's standardized math scores by subtopic

Dependent variable:	Immediate effect (in σ)		Effect after one year (in σ)	
Subject domain of items:	NSEA	GEOM & DSP	NSEA	GEOM & DSP
	(1)	(2)	(3)	(4)
Treatment	0.25***	0.30***	0.09	0.06
	(0.09)	(0.09)	(0.11)	(0.11)
Baseline score	0.63***	0.49***	0.42***	0.48***
	(0.10)	(0.11)	(0.13)	(0.11)
Adjusted R^2	0.73	0.72	0.60	0.67
Observations	164	164	136	136
Teacher controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes

Notes: The *immediate effect* is estimated based on the endline data collected in Sept. 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in September 2020. *NSEA:* Items covering number sense and elementary arithmetics; *GEOM & DSP:* Items covering geometry and measurement as well as data, statistics, and probability. In all columns, the share of correct answers (by subject domain) is standardized to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Huber-White robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

A.5.1 Program Effects by Grade Level

Table A.4: ITT-Estimates on the effects on teacher's math scores by grade level of items

Dependent variable:	Immediate effect (in σ)				Effect after one year (in σ)			
Grade level of items:	Gr. 2/3	Gr. 4	Gr. 5	Gr. 6	Gr. 2/3	Gr. 4	Gr. 5	Gr. 6
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.19	0.19*	0.31***	0.36***	0.12	−0.09	0.09	0.16
	(0.12)	(0.10)	(0.10)	(0.11)	(0.14)	(0.12)	(0.12)	(0.11)
Baseline score	0.35***	0.08	0.58***	0.38***	0.29**	0.15	0.52***	0.36***
	(0.09)	(0.10)	(0.11)	(0.09)	(0.12)	(0.12)	(0.14)	(0.11)
Adjusted R^2	0.58	0.63	0.68	0.60	0.38	0.56	0.54	0.64
Observations	164	164	164	164	136	136	136	136
Teacher controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *immediate effect* is estimated based on the endline data collected in Sept. 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in Sept. 2020. In all columns, the share of correct answers (by grade level of items) is standardized to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Huber-White robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

A.6 Program Effects by Teachers' Baseline Ability and Age

We estimate the following regression equation:

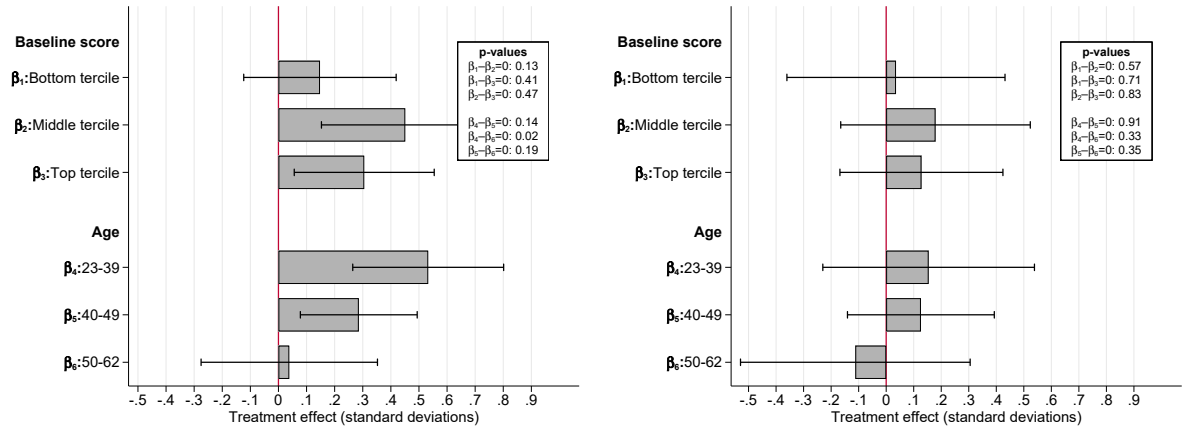
$$Y_{jk}^{EL} = \alpha + \beta\, Treat_{jk} + \lambda(Treat_{jk} \times Covariate_{jk}) + \delta Y_{jk}^{BL} + X_{jk}^{'}\gamma + S_{jk}^{'}\rho + \phi_k + \epsilon_{jk}, \quad \text{(A.1)}$$

where $Treat_{jk} * Covariate_{jk}$ denotes the interaction of the treatment dummy and the specific variable of interest (i.e. teacher baseline score, gender, age). The coefficient λ then captures the extent to which the effect of the treatment differs along these interacted characteristics. All other terms are defined as in Equation (1).

Table A.5: Effect heterogeneity along baseline ability and age

Dependent variable:	Immediate effect (in σ)		Effect after one year (in σ)	
Covariates:	Baseline score	Age	Baseline score	Age
	(1)	(2)	(3)	(4)
Treatment	0.29***	0.28***	0.08	0.07
	(0.08)	(0.08)	(0.10)	(0.10)
Covariate	0.85***	−0.01*	0.67***	−0.02
	(0.11)	(0.01)	(0.14)	(0.01)
Treatment x covariate	0.03	−0.02*	0.03	−0.02
	(0.07)	(0.01)	(0.10)	(0.01)
Adjusted R^2	0.80	0.81	0.71	0.72
Observations	164	164	136	136
Baseline math score	Yes	Yes	Yes	Yes
Teacher controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes

Notes: The *immediate effect* is estimated based on the endline data collected in Sept. 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in Sept. 2020. In all columns, the share of correct answers is standardized to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Huber-White robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.



(a) Immediate program effects (in σ)

(b) Program effects after one year (in σ)

Figure A.3: Effect heterogeneity by baseline score and and age
Note: Same set of controls as in Table A.7. Spikes show 95% confidence intervals.

A.7 Program Effects Estimated with Fully Balanced Sample

Table A.6: ITT-estimates for the program effects on teachers' math scores with constant sample

	Immediate effect				Effect after one year			
Dependent variable:	Percent correct		Standardized		Percent correct		Standardized	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	5.21***	5.28***	0.27***	0.28***	0.71	1.58	0.04	0.09
	(1.61)	(1.66)	(0.08)	(0.09)	(1.83)	(1.80)	(0.10)	(0.10)
Baseline score	0.83***	0.73***	0.84***	0.75***	0.76***	0.60***	0.81***	0.64***
	(0.09)	(0.11)	(0.09)	(0.11)	(0.12)	(0.13)	(0.13)	(0.14)
Adjusted R^2	0.80	0.81	0.80	0.81	0.70	0.73	0.70	0.73
Observations	131	131	131	131	131	131	131	131
Teacher controls	No	Yes	No	Yes	No	Yes	No	Yes
School controls	No	Yes	No	Yes	No	Yes	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *immediate effect* is estimated based on the endline data collected in September 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in September 2020. In *columns (3), (4), (7), and (8)*, the share of correct answers is standardized to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Huber-White robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

A.8 Program Effects Estimated with Entropy Balancing

Table A.7: ITT-estimates for the program effects on teachers using entropy balancing

	Immediate effect				Effect after one year			
Dependent variable:	Percent correct		Standardized		Percent correct		Standardized	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	5.33***	5.33***	0.28***	0.28***	1.35	1.35	0.07	0.07
	(1.46)	(1.46)	(0.08)	(0.08)	(1.77)	(1.77)	(0.10)	(0.10)
Baseline score	0.88***	0.85***	0.89***	0.86***	0.74***	0.67***	0.79***	0.71***
	(0.09)	(0.10)	(0.09)	(0.10)	(0.12)	(0.13)	(0.13)	(0.14)
Observations (weighted)	175	175	175	175	175	175	175	175
Observations (unweighted)	164	164	164	164	136	136	136	136
Teacher controls	No	Yes	No	Yes	No	Yes	No	Yes
School controls	No	Yes	No	Yes	No	Yes	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *immediate effect* is estimated based on the endline data collected in September 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in September 2020. In *columns (3), (4), (7), and (8)*, the share of correct answers is standardized to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Robust standard errors accounting for uncertainty induced by entropy balancing in parentheses, see Jann (2021) for methodological details. * p<0.10, ** p<0.05, *** p<0.01.

A.9 The Effect of Teacher Content Knowledge on Student Learning

This section comprehensively discusses evidence on the effect of teacher content knowledge on student learning: *First*, we present estimates based on Salvadoran data. *Second,* we summarize international evidence from Asia, Africa, and South America. *Third,* we discuss a possible quantification in terms of school year equivalents.

Estimates based on Salvadoran Data. Our data on teacher content knowledge can be combined with data on student learning outcomes collected during a field experiment in 2018 (for details see Büchel et al., 2021).[9] The teacher survey was administered towards the end of the school year 2018, which also marked the end of the aforementioned experiment. However, as the assignment of teachers to classes was not experimentally manipulated, we do not claim that the reported correlations are causal. In line with standard practice, we specify the basic model for estimating the relation between teacher content knowledge and students' learning as

$$\Delta\tilde{Y}_i = \alpha + \beta\tilde{S}_j + G_i\gamma + T_i\delta + \epsilon_i.$$

The dependent variable, $\Delta\tilde{Y}_i$, is student i's learning defined as $\Delta\tilde{Y}_i = \tilde{Y}_i^2 - \tilde{Y}_i^1$ with $\tilde{Y}_i^1 = (Y_i^1 - \bar{Y}^1)/\hat{\sigma}_{Y^1}$ and $\tilde{Y}_i^2 = (Y_i^2 - \bar{Y}^1)/\hat{\sigma}_{Y^1}$, where Y_i^1 and Y_i^2 are the student's IRT scores in wave 1 and wave 2, respectively, and $\bar{Y}^1$ and $\hat{\sigma}_{Y^1}$ are the mean and standard deviation of the scores in wave 1. The predictor of interest is $\tilde{S}_j$, the standardized knowledge score of teacher j (who teaches student i), defined as $\tilde{S}_j = (S_j - \bar{S})/\hat{\sigma}_S$ where S_j is the percentage of correct answers that teacher j achieved in the assessment. The model further includes an indicator vector for the student's grade, G_i, since teacher knowledge is correlated with grade and ability improvements are smaller among higher-grade students. Furthermore, the treatments imposed as part of the field experiment did affect learning so that teacher effects are evaluated within treatment groups as captured by the indicator vector T_i.

This basic model corresponds to specification (1) in Table A.8. Additional specifications include class-level controls (columns 2–4), school-level controls (columns 3 & 4), and additional teacher characteristics (column 4).[10] All specifications yield a positive relation between teacher content knowledge and student learning. Quantitatively, a 1σ increase in teacher knowledge is associated with a 0.09σ to 0.12σ gain in student learning.

International Evidence. In Table A.9, we compare our estimates reported in Table A.8 to recent evidence reported for primary schools in Peru (Metzler and Woessmann, 2012), Africa (Bietenbeck et al., 2018; Bold et al., 2019), and Pakistan (Bau and Das, 2020). While the evidence unambiguously demonstrates that better content knowledge of teachers improves student learning, the effect magnitude varies by subject. Studies distinguishing between math and language find that teachers' content knowledge plays a more decisive role in the instruction of math. Evidence for Peru, Pakistan and El Salvador consistently suggest that a 1σ increase in teacher content knowledge is associated with an annual gain in students' math scores of about 0.09σ. With respect to language, less evidence is available and the correlation is weaker. The estimated coefficients vary between 0.03 (insig.) and 0.06 for languages, and between 0.03 and 0.06 when the effect of teacher content knowledge is estimated across multiple subjects.

[9] The analysis draws on the same sample of students as the cited field experiment, except for the following limitations: Four teachers did not attend the assessment so that we drop their classes from the sample. We also eliminate five classes that were re-assigned to a new teacher during the school year 2018. Finally, we only include teachers who provided information on all covariates; this excludes another eleven classes from the sample.

[10] *Teacher controls* comprise age, sex, highest degree, experience as a math teacher, and travel time to school. *Class level controls* are class size, sex ratio, avg. household size, avg. household wealth, avg. maternal literacy rate within the class, and a binary indicator for afternoon classes. *School level controls* encompass an infrastructure index, an equipment index, travel time to the department's capital as well as binary indicators for student access to a computer lab, exposure to gangs, and location in a rural area.

Table A.8: Relation between teacher's test score and students' learning over an eight month evaluation period and in a sample of Salvadorian primary school classes of grades 3 to 6.

	Student learning gains							
	Standardized (σ)				School year equivalents			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standarized teacher score	0.098***	0.091***	0.097***	0.116***	0.276***	0.256***	0.274***	0.324***
	(0.031)	(0.032)	(0.031)	(0.036)	(0.083)	(0.085)	(0.080)	(0.091)
Grade level fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Class level controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
School level controls	No	No	Yes	Yes	No	No	Yes	Yes
Teacher controls	No	No	No	Yes	No	No	No	Yes

Notes: As the student data was collected for an experimental evaluation of a computer-assisted learning intervention, all models control for the treatment assignment of classes. Number of observations: 2786 students, 120 teachers, 48 schools. School-level clustered standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

The finding that content knowledge of teachers has a stronger impact on learning outcomes in math is consistent with studies from OECD countries reporting greater variance in teacher effects on achievement in math than language. One reason may be that math is almost exclusively learned in the classroom, while languages are learned to a great extent outside of school (e.g. Jackson et al., 2014).

Converting Salvadoran Estimates to School Year Equivalents. To assess the magnitude of the relation between teachers' content knowledge and student learning, it is informative to express learning gains in school year equivalents. To do so, we use our Salvadoran data introduced above and compute each student's difference in IRT scores between wave 1 and 2 and divide it by the average score difference between grades, so that results are expressed in units of children's average learning gains during one school year. Formally, we replace the dependent variable $\Delta\tilde{Y}_i$ with

$$\Delta Y_i^E = (Y_i^2 - Y_i^1)/\hat{\gamma}$$

where $\hat{\gamma}$ is the slope coefficient of student's grade $\tilde{G}_i$ in model

$$Y_i = \alpha + \gamma\tilde{G}_i + T_i\delta + \epsilon_i$$

estimated using data from wave 2. Treatment indicator vector T_i is included in the model to eliminate a biasing effect of the CAL intervention that took place between wave 1 and wave 2.

Replicating columns (1) to (4) in Table A.8 with student learning measured in school year equivalents suggests that a 0.3σ increase in a teacher's math score is associated with 0.08 to 0.1 additional years of schooling (see columns 5 to 8 in Table A.8). Accordingly, shifting a student from a teacher at the lowest to one at the highest decile would yield 0.7 to 0.9 additional years of schooling, and hence almost double the students' annual progress in math.

Table A.9: Evidence on the effect of teacher content knowledge on student learning in developing countries

	Metzler & Wössmann (2012)	Bietenbeck et al. (2018)	Bold et al. (2019)	Bau & Das (2020)	This study's results, Table A.8
Main effect (per year)					
$+1\sigma$ teacher test score on student test scores (in σ)	*Math:* 0.09 *Lang.:* 0.03 (insig.)	*Mixed:* 0.03	*Mixed:* 0.07	*Math:* 0.09 *Language:* 0.06	*Math:* 0.09–0.12
Sample					
Country and region	Peru	6 East African countries	7 African countries	Pakistan, Punjab	El Salvador, Morazán
Subjects	Math Language	*Mixed*: Math and language	*Mixed*: Math and language	Math Language	Math
Level of education	Primary school (Grade 6)	Primary school (Grade 6)	Primary school (Grade 4)	Primary school (Grades 3–5)	Primary school (Grades 3–6)
Empirical strategy	Teacher FE + Student FE	Teacher FE + Student FE	Teacher FE + Student FE	Teacher value-added approach	various controls

Sources for estimates reported in first row: Metzler and Woessmann (2012): Table 2, column 1. Bietenbeck et al. (2018): Table 3, column 5. Bold et al. (2019): Table 4, column 3. Bau and Das (2020): Table 3 (columns 2–6), Table 4 (column 7).

B Appendix: Methods

B.1 Sampling for the Representative Teacher Assessment

Our base population encompasses all primary school math teachers teaching at least one class between grades 3 and 6 in one of the 306 public primary schools in the department of Morazán, El Salvador. Ten out of the 306 public schools in Morazán registered zero students in these grades and were excluded, leaving 296 schools in the population. Since the teacher assessment took place in the context of a randomized controlled trial on a computer assisted learning (CAL) project (Büchel et al., 2021), our sample is drawn from two strata of schools.

1. *Schools that were eligible for the CAL project*: Of the 296 public primary schools with classes in grades 3 to 6 in Morazán, 57 schools fulfilled the eligibility criteria for the CAL project (defined in terms of school size, security situation, accessibility, and electrification). In these 57 schools, 198 classes from grades three to six were randomly chosen to be part of the CAL experiment. All math teachers instructing at least one of these classes are also included in the target sample of the present study. Teachers from this stratum of schools had a probability of 65.7% of becoming part of our sample and are thus over-sampled relative to the base population.

2. *Schools that were not eligible for the CAL project*: Among the remaining 239 schools, 50 schools were randomly selected, stratified by 16 geographical regions, and all math teachers in grades 3 to 6 in these schools were invited to participate in the assessment. Teachers from this stratum of schools had a sampling probability of 21%.

In our data analyses, we take account of the described stratification, the unequal sampling probabilities, as well as the fact that schools, not teachers, are the primary sampling unit (using Taylor-linearization for variance estimation).

B.2 Sampling and Randomization for the Field Experiment

As illustrated in Figure B.1, the sampling and randomization procedure for the field experiment consisted of five steps.

1. All public primary schools with students in grades three to six in Morazán serve as the starting point for the sampling process.

2. For implementation purposes, the 49 smallest schools with fewer than a total of 15 students in grades one to six were excluded, resulting in a target population of 253 schools.

3. The NGO sent out invitations to all grade three to six math teachers in eligible schools. Overall, 313 teachers from 175 different schools applied for the program.

4. Before the start of the intervention, all candidates took an unannounced baseline assessment.[11] Based on the results of this assessment, the worst-performing applicant of every school was selected for participation. Note, however, that this part of the sampling procedure was not communicated to applicants to avoid misaligned incentives during the assessments. At the end of this procedure 175 teachers from 175 different schools across Morazán remained in the sample.

[11] A sub-group of the applicants took the math assessment in the context of the representative math assessment (see section B.1). In March 2019, the same assessment was administered to all other applicants. The proportion of teachers who took the exam in September 2018 (instead of March 2019) does not differ significantly between the control and the treatment group. In both cases, the assessment was unannounced.

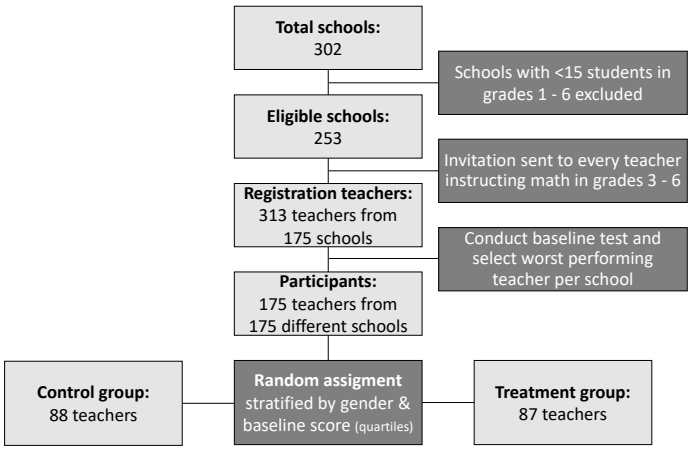


Figure B.1: Sampling and randomization scheme

5. In a final step step, the 175 pre-selected teachers were randomly assigned to either the control group (88 teachers) or the treatment group (87 teachers). To enhance the efficiency of the estimates, randomization was stratified by the teachers' baseline score and gender. For this purpose, teachers were grouped by performance quartiles using the baseline assessment and by gender so that we obtained eight strata. Even though we randomized at the teacher level, the pre-selection left only one teacher per school in the sample. This prevents potentially biased estimates due to spillover effects within schools.

B.3 Assessment Design

To design the math tests for the representative teacher assessment and the three assessments for the field experiment, we proceeded as follows.

1. We first summarized the Salvadoran math curriculum for grades two to six along the three topics *Number Sense & Elementary Arithmetic* (NSEA), *Geometry & Measurement* (GEOM), and *Data, Statistics & Probability* (DSP).

2. For the assessments, we then mapped test items from various sources on the Salvadoran curriculum. These sources include official textbooks of El Salvador, publicly available items from the STAR evaluations in California, publicly available items from the VERA evaluations in Germany, and publicly available items from the SAT assessments in Britain.[12]

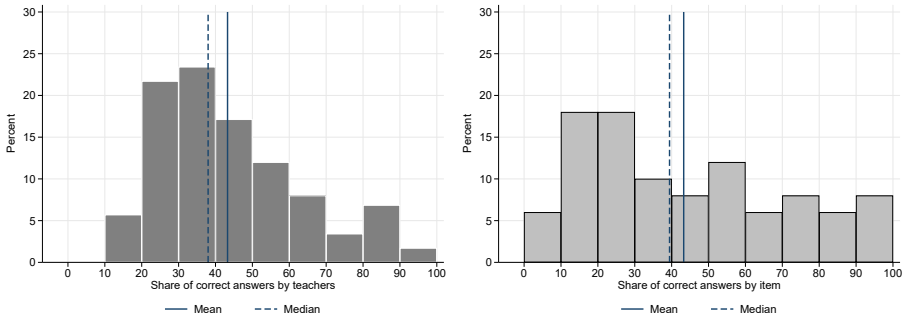
3. We then designed paper and pencil math assessments including a total of 50 questions on materials from grade two (∼6 items) and grades three to six (between 10 and 13 items) reflecting the official national curriculum. The assessments cover questions from NSEA (∼30 items), GEOM (∼15 items), and DSP (∼5 items) and are meant to be completed in 90 minutes. The relative weighting of the three main domains emulates the weighting in the national primary school math curriculum. To make sure that questions are suitable for the Salvadoran context, assessments were reviewed by local teaching experts and the local education ministry. Moreover, the exam lasted a generous 90 minutes to guarantee that every participant had enough time to carefully draft the answers so that wrong answered cannot be attributed to time pressure.

[12]Further information on the *Standardized Testing and Reporting (STAR)* program in California is available online: `www.cde.ca.gov/re/pr/star.asp`. *VERA* is coordinated by the Institut für Qualitätsentwicklung im Bildungswesen (IQB), see `www.iqb.hu-berlin.de/vera`. *SAT* is an acronym for *standardized assessment tests* coordinated by the UK's Standards and Testing Agency, see `https://www.gov.uk/government/organisations/standards-and-testing-agency`.

4. Based on these assessments, we used two different main outcome measures at the teacher level: the share of correctly answered questions and standardized test scores. All results in the field experiment are based on double coded data by pre-trained staff in El Salvador (batch 1) and Switzerland (batch 2 plus harmonization of batches 1 and 2).

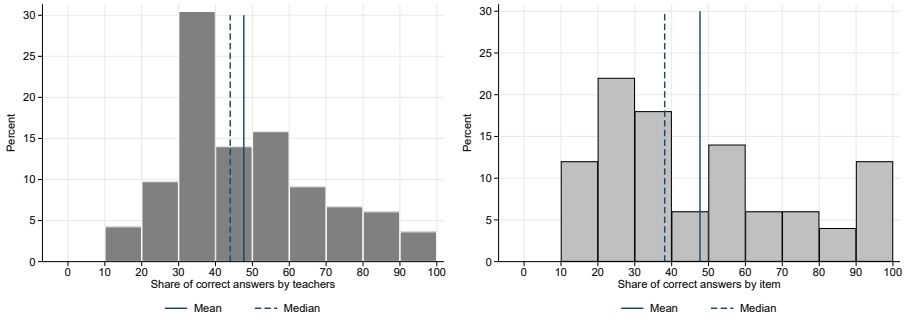
B.4 Assessment Diagnostics

Figure B.2 presents the distribution of correct answers by teachers and by items for the baseline, endline and follow-up assessments. The histograms show that there are neither floor nor ceiling effects. Teachers were able to answer at least 10 percent of the items in the baseline, 16 percent in the endline and 18 percent in the follow-up assessment. On the other hand, no teacher scored 100 percent correct answers in any of the waves. Further, there is no item that was not answered correctly by anyone (minimum share of correct answers across all waves and items is 3 percent) or an item which was solved successfully by all teachers (maximum share of correct answers is 98 percent across all waves and items).



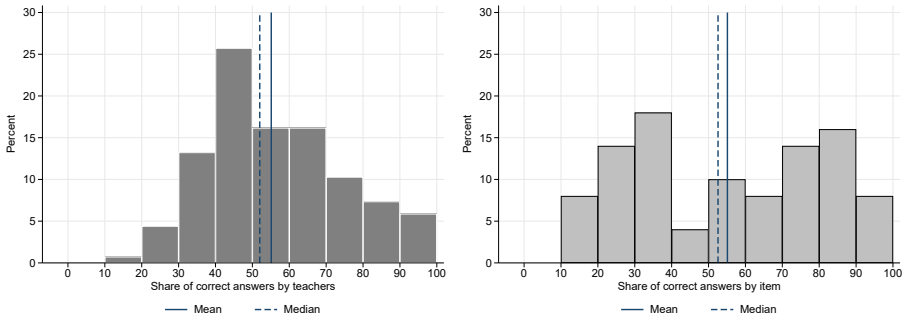
(a) Baseline assessment by teachers

(b) Baseline assessment by items



(c) Endline assessment by teachers

(d) Endline assessment by items



(e) Follow-up assessment by teachers

(f) Follow-up assessment by items

Figure B.2: Share of correct answers across items and teachers by assessment