When Better Teachers Aren't Enough:

An Experimental Evaluation of Teacher Training Programs in El Salvador

Carla Coccia*, Martina Jakob†, Konstantin Büchel‡, Ben Jann§

January 30, 2026

Despite billions spent annually on teacher training, rigorous evidence on standalone in-service programs remains scarce, as most evaluated programs bundle training with curriculum or material reforms. We address this gap through a large-scale randomized controlled trial with 338 schools and over 6,000 students in El Salvador. Teachers are randomly assigned to either a control group or one of three training programs focusing on (i) pedagogical knowledge, (ii) content knowledge, or (iii) a combination of both inputs. We find lasting effects on teacher content and pedagogical knowledge of up to 0.3 and 0.5 σ respectively one year after program end. Yet, this only changes teachers' classroom practices in the short-run and does not translate into significant student learnings. The data most closely aligns with a setting where teachers face a dual challenge: introducing new ideas in a rigid environment while navigating the significant learning gaps present among students in later grades.

*The project is funded by the Swiss National Science Foundation under grant number 197576 for a total of 750,000 CHF, and implemented by Fundación Consciente El Salvador with approval from the Ministry of Education of El Salvador. This project received IRB approval from the Ethics Committee of the Faculty of Business Administration, Economics and Social Sciences of the University of Bern on November 23, 2020. It was registered in the AEA Registry on September 21, 2022 (AEARCTR-0010035). We thank Tessa Bold, Jacobus Cilliers, Mauricio Romero, Lennart Schalk, Caroline Strobl and Romana Tedeschi helpful advice on the project. We thank Stefanie Clausen, Benedikt Finger, Salvador Reyes, Alonzo Solis and Isabel Gonzales for excellent implementation design and coordination. We thank Jenny Mosimann and Elena Meier for excellent research assistance.

*University of Bern; E-mail: carla.coccia@unibe.ch

†University of Zurich; E-mail: martina.jakob@econ.uzh.ch

‡Youth Impact; E-mail: kbuchel@youth-impact.org

§University of Bern; E-mail: ben.jann@unibe.ch

1 Introduction

Quality education remains elusive in many developing countries, even as school enrollment has expanded substantially in recent decades. International assessments consistently show that the rate at which education systems convert schooling into human capital is low. This has prompted the World Bank to dedicate its 2018 World Development Report to what was declared a global ”learning crisis” (Filmer and Rogers, 2019). Evidence from Africa, Asia, and Latin America indicates that shortcomings in teachers' pedagogical and content knowledge are a major barrier to more effective schooling systems, implying that the learning crisis is rooted in a deeper teaching crisis (Bold et al., 2017a). Without joint efforts, this situation is likely to reproduce itself: Many of today's poorly qualified teachers will continue teaching for years to come and consequently shape tomorrow's teachers.

Improving teacher quality is therefore a stated policy priority of many governments in developing countries. Each year they allocate billions of dollars and personnel-hours to teacher professional development programs (Loyalka et al., 2019). Also large donors invest at scale: nearly two-thirds of World Bank education projects between 2000 and 2012 included teacher professional development components (Popova et al., 2018). Nonetheless, the evidence base on which program designs improve teacher practice remains surprisingly thin (Popova et al., 2022). Existing evaluations tend to focus on programs bundled with broader curricular reforms, new instructional materials or technological changes, leaving a limited understanding of the conventional in-service training models that dominate public-sector provision.

We adress these gaps through a large-scale randomized controlled trial (RCT) across 338 schools in eastern El Salvador, a setting characterized by low levels of both teacher pedagogical and content knowledge. We examine whether well-implemented, standalone in service training can improve student learning, and whether impacts differ by an emphasis on pedagogical versus content knowledge. We randomly assign 338 upper-primary math teachers to either a control group or one of three training programs focusing on (i) pedagogical knowledge, (ii) content knowledge, or (iii) a combination of both inputs. All training programs share a common basic framework combining face-to-face workshops, coaching elements, and self-study modules. The program was developed in collaboration with educational experts from the Swiss University of Teacher Education Fribourg and Consciente, a Salvadorian NGO specializing in evidence-based schooling projects. It was approved and supported by the Salvadorian Ministry of Education.

To evaluate the effects of the interventions and examine the mechanisms through which they operate, we assemble rich data on teacher knowledge, instructional behavior, and student achievement over two school years. Outcomes are measured at baseline, immediately after program completion, and one year later. Teacher content knowledge is assessed using math tests that span the full primary math curriculum, with particular emphasis on higher-grade material. Pedagogical knowledge is measured at follow-up using two complementary tasks: lesson plan design for a specified topic and classroom context and responses to a classroom management scenario. Instructional practices are measured at endline and follow-up through classroom observations based on a modified Stallings protocol (Stallings, 1973), which records instructional time allocation and includes additional low-inference pedagogical items. Finally, student achievement is captured using curriculum-aligned math tests administered to one fourth-grade and one fifth-grade classroom per teacher.

Baseline descriptive evidence reveals severe shortfalls in math knowledge among both students and teachers. On average, fourth and fifth graders answer 72% of first-grade items correctly, but accuracy drops by more than half—to 35%—for second-grade items and further to 26% and 12% for third- and fourth-grade items respectively. Teachers likewise demonstrate incomplete mastery of the primary curriculum, correctly answering just 61 percent of grade two to six items on average, in line with existing evidence from El Salvador (Brunetti et al., 2024). To benchmark these outcomes internationally, we map student and teacher performance onto the TIMSS scale using Item Response Theory. Salvadorian students rank near the bottom of the global distribution, second only to the lowest-performing country in the TIMSS sample. Notably, the average Salvadorian teacher barely outperforms the average fourth-grade student in Singapore, the top-performing system in TIMSS, underscoring the depth of content knowledge gaps in the education system.

To benchmark ex ante expectations, we elicited predictions from education researchers and practitioners on the social science prediction platform. Based on a detailed description of the intervention and its theory of change, respondents predicted sizable improvements in teacher knowledge, classroom practices, as well as student learning across all treatment arms. Mean predicted effects for student achievement were around 0.15 standard deviations, marginally smaller than expected teacher gains (∼0.16–0.17 sd) but still substantial (see Figure A11). Only four out of 38 respondents anticipated student effects below 0.05 standard deviations, and among them, just one predicted negligible student gains while expecting large improvements for teachers. These priors suggest that most informed observers expected teacher trainings to improve student learning, and that the few skeptical views were driven by doubts about teacher learning rather than a failure of knowledge transmission.

In line with expectations, we find substantial effects on teachers for all three programs. The math program increases teachers' content knowledge by 0.14σ at endline and 0.29σ one year later, indicating consolidation of knowledge over time. The combined program yields smaller but also growing gains (0.11σ at endline and 0.15σ at follow-up). Consistent with its design, the pedagogy-only program produces no measurable improvement in teachers' math knowledge. Across all three arms, however, we estimate large improvements in pedagogical knowledge at follow-up, ranging from 0.47σ to 0.57σ. Despite these changes in teacher competencies, we detect no gains in student achievement immediately after implementation or one year later; estimates are small with standard errors that rule out moderate effects.

Drawing on our complementary classroom observations, teacher survey data and semi-structured interviews, we examine four key steps linking teacher trainings to student learning, as outlined in our theory of change: (i) successful implementation and high compliance, (ii) meaningful improvements in teacher knowledge, (iii) translation of these gains into classroom practice, and (iv) students' ability to absorb and retain instruction. The evidence points to two central constraints. First, institutional rigidity, particularly reliance on standardized curricula and scripted textbooks, appears to limit the sustained adoption of newly acquired teaching strategies: while observed pedagogical practices improve at endline ($\approx 0.35\sigma$), these gains dissipate within a year. Second, and perhaps most critically, a sharp misalignment between the prescribed curriculum and students' actual learning levels point to a major bottleneck to student progress. Polynomial estimates show positive treatment effects only for students at or above grade level at baseline, while effects are statistically indistinguishable from zero for students further below.

This study's contribution to the literature is fourfold. First, we clarify an important but often overlooked distinction in the literature on teacher-training interventions: standard in-service training differs from comprehensive structured pedagogy reforms. The most influential positive results in the teacher-training literature come from comprehensive interventions that simultaneously alter what teachers teach and how they teach—typically bundling training with revised curricula, structured materials, and ongoing coaching (see for example Kerwin and Thornton, 2021; Buhl-Wiggers et al., 2023; Cilliers et al., 2020; Bruns et al., 2018; Bassi et al., 2020; Albornoz et al., 2020; Piper et al., 2018; Jukes et al., 2017). Our review of two recent and comprehensive meta-studies (Popova et al., 2022; Angrist et al., 2025)—supplemented by a keyword search ("teacher training," "professional development") of the Social Science Registry—reveal a substantive empirical gap: Traditional in-service models that dominate public systems continue to absorb considerable resources but remain under-evaluated in their conventional form.[2] Existing evidence on stand-alone teacher training largely focuses on early childhood education (Özler et al., 2018; Yoshikawa et al., 2015; Wolf, 2018) or on targeted pedagogical principles such as Teaching at the Right Level (Banerjee et al., 2016), curiosity-centered instruction (Alan and Mumcu, 2024) or learning-to-learn methods (Ashraf et al., 2020). Only two recent experimental studies assess classical government training programs in primary schools, and both find zero effects, largely due to either low uptake or weak implementation quality as is common in government settings (Loyalka et al., 2019; Schaffner et al., 2025).[3] We study a well-implemented, NGO-led program intentionally designed to overcome these weaknesses.It produces sizable improvements in teacher content knowledge, pedagogical knowledge, and short-term instructional practices. Yet student learning outcomes remain unchanged. By unpacking the mechanisms underlying this disconnect, we contribute to understanding why enhanced teaching performance does not necessarily lead to better learning.

Second, this study contributes to filling a considerable evidence gap identified by the literature regarding content-based teacher training in developing countries (Bold et al., 2017b). Despite teachers' deficits in content knowledge documented in many developing settings (Bold et al., 2017b; Sinha et al., 2016; Brunetti et al., 2024), most teacher training interventions focus on improving instructional practices rather than content mastery. This is notable given non-experimental evidence showing that observed differences in teachers' content knowledge account for roughly 30 percent of students' learning shortfalls relative to the curriculum and about 20 percent of cross-country achievement gaps (Bold et al., 2019). Studies also indicate that content knowledge plays a more decisive role in mathematics instruction than in language (Metzler and Woessmann, 2012; Bold et al., 2019). Some evaluated programs complement pedagogy with measures to improve teachers' content knowledge, but very few, if any, put significant time and effort into content-based training. Reflecting this state

[2] (Angrist et al., 2025) is among the first to formalize this distinction by explicitly separating structured-pedagogy reforms from other interventions involving the training of teachers, underscoring its emerging relevance for interpretation of the evidence base.

[3] A third, related study from Mexico evaluates an NGO-implemented training program for secondary-school English teachers and reports improvements in teacher knowledge and classroom behavior, but the design's reliance on a subject-specific shift to English-only instruction, together with limited statistical power, makes it difficult to compare to general teacher-training reforms (Bando and Li, 2014).

of the literature, Bold et al. (2017b, p. 202) conclude that "[u]nfortunately, there are few, if any, well-identified studies on how to effectively improve teacher knowledge and skills and the impact thereof". We contribute to filling this gap by isolating the effects of content- versus pedagogy-based training and testing complementarities between these approaches. To our knowledge, no study in a developing-country setting has yet systematically done so.

Third, we contribute to recent efforts and calls by the literature to improve the comparability of impact estimates across educational interventions (Angrist et al., 2025). Over the past decade, a growing body of meta-studies has sought to synthesize and generalize evidence from educational trials (e.g., Kremer et al., 2013; McEwan, 2015; Glewwe, 2016; Ganimian and Murnane, 2016; Evans and Yuan, 2022; Angrist et al., 2025). Yet the dominant practice of reporting effect sizes in standard deviations, or in measures derived from them, limits the interpretability of such syntheses. Standard deviations are sample dependent, becoming smaller in more heterogeneous populations, and they are test dependent, becoming larger when assessments closely mirror program content. Following Patel and Sandefur (2020), we address these limitations by constructing IRT-based assessments and linking them to an international benchmark, allowing impacts to be reported on the TIMSS scale. This results in effect-size metrics that are conceptually invariant to sample composition and test design, thereby improving comparability across interventions, populations.

Fourth and last, our paper ties into the recent literature systematically documenting learning levels across the world (Angrist et al., 2021; Patel and Sandefur, 2020). By benchmarking our assessments to the TIMSS test, we place students' performance on a metric that is directly comparable across countries. To our knowledge, we are the first to implement Patel and Sandefur (2020)'s approach in a field experiment, demonstrating how research on educational interventions can simultaneously generate data with global interpretability. This approach yields a de facto global standardized scale without requiring children to sit for the same exam. In addition, we extend their framework by accounting for differential item functioning (DIF) in the mapping of scores to the TIMSS distribution and by relaxing the assumption that mapped scores follow a normal distribution.

2 Research Design

2.1 Context and Interventions

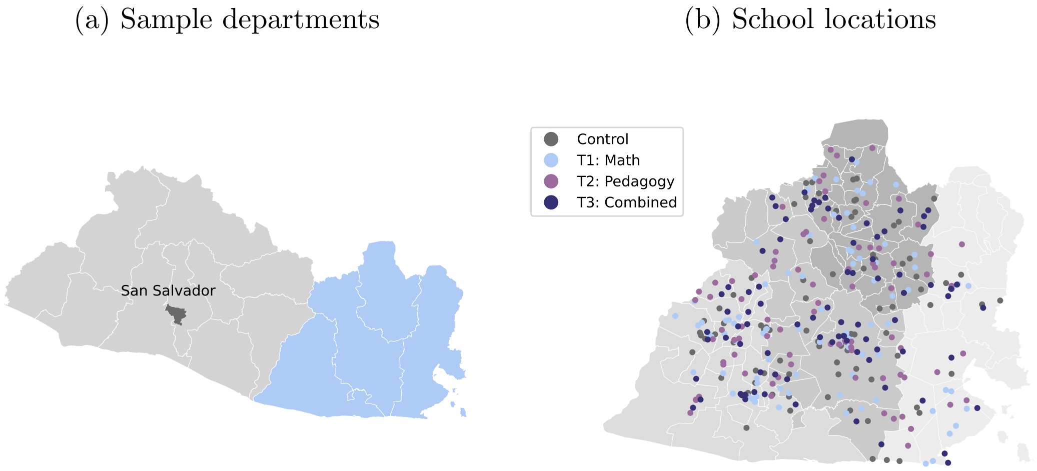
Our study is set in El Salvador, a lower-middle-income country in Central America. It was conducted in its oriental region which comprises 4 out of the 14 existing departments of the country: Morazán, Usulután, San Miguel and La Unión (see Figure 1). El Salvador's education paints a picture no different from that of many developing countries: With some children unable to read or write at the end of primary school, students' learning outcomes fail to meet even the most basic standards. In fact, recent World Bank data shows that the quality of primary education in El Salvador fall below the average of lower middle-income countries (Angrist et al., 2021).

A prominent feature of El Salvador's education system is the low level of teachers' content knowledge and pedagogical skills. Teachers lack a basic mathematical understanding of the very concepts they are required to teach. Correspondingly, a recent study by Brunetti et al. (2024) finds that the average, representative teacher in Morazán, one of the study departments, answers more than half of grade two to six questions incorrectly. The study concludes, that a mere 14 percent of teachers are actually equipped with the necessary content knowledge to effectively teach mathematics at the primary school level [4]. Despite having completed 13 to 17 years of formal education, the majority of primary school teachers are therefore "confronted with the daunting task of teaching what they don't know" (Brunetti et al., 2024, p. 209).

At the same time, anecdotal evidence also points to major shortfalls in teachers' pedagogical knowledge. Practical pedagogy training does not form part of the official curriculum of El Salvador's national teacher training institutions in El Salvador. Consequences are evident: The traditional chalk-and-talk teaching style, centered almost exclusively on the blackboard and the teacher's exposition, continues to dominate Salvadorian classrooms. As a result, teachers tend to be more focused on adhering to and covering the school curriculum as determined by the government text book, rather than on the actual learning progress of their students.

In this context, we partnered with the local NGO Consciente and experts from the Swiss University of Teacher Education Fribourg with the aim of strengthening teacher skills and enhancing student learning outcomes in public primary schools. We implemented a large-scale randomized control trial that compares three compre-

[4] Based on the minimum proficiency threshold proposed by (Bold et al., 2017a).


(a) Sample departments

(b) School locations

Figure 1: Study Area

hensive teacher training programs, each targeting a distinct area: (i) mathematical content knowledge (ii) pedagogical skills and (iii) a combined approach.

Each intervention was delivered over a six-month period. The training structure was divided into seven three-week blocks that include a mix of in-person workshops, self-study, and personalized coaching (see Figure 2). The program and program materials were developed in a joint effort of Swiss and Salvadorian education experts and approved as well as supported by the Salvadorian Ministry of Education. The content of the training blocks for mathematics was designed to prioritize the most relevant aspects of teaching upper primary grades while ensuring a foundational review of key mathematical principles. For pedagogy, it covered the essential didactic skills required to plan, pace, and manage instruction, and to adapt teaching to students' learning levels. The combined approach (iii) was streamlined to focus on the most impactful components of both the math and pedagogical interventions (a detailed overview of the covered topics can be found in Figure A1).[5]

The trainings were explicitly designed to satisfy rigorous pedagogical standards and featured substantial interactive elements, including clear learning objectives, structured practice with feedback and collaborative problem-solving activities. To facilitate transfer to teachers' day-to-day instruction, all trainings followed the same instructional sequence used in El Salvador's nationally mandated ESMATE mathematics textbooks: introduction, content, practice, summary. The mathematics training followed the same design logic as the other two trainings, thereby enabling teach-

[5] Full materials available here: https://mdid.consciente.ong

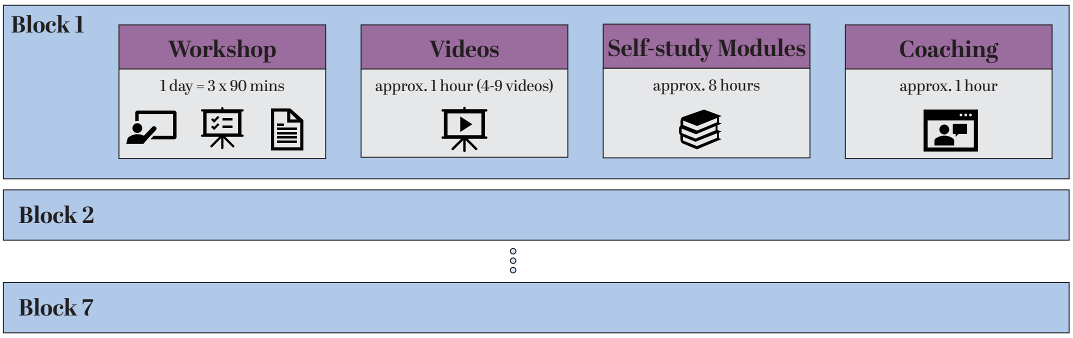


Figure 2: Program structure

Notes: Each training block has a duration of three weeks. In-person workshops are thus three weeks apart.

ers not only to strengthen their content knowledge but also to benefit from more effective pedagogical strategies for conveying key concepts. Following the evidence in Popova et al. (2022) on most impactful training practices, the program incorporated structured lesson enactment as well as extensive in-person sessions. Additionally, in line with the same evidence, completion of the training was tied to career incentives through Ministry of Education certification.

The teacher trainings were conducted by 10 teaching professionals employed by the partnering NGO, Consciente. Prior to delivering the training, these instructors underwent an extensive training phase lasting several weeks, during which they were thoroughly acquainted with the training content and materials.[6] The training for the NGO-hired instructors was directly conducted by the Swiss-Salvadorian team that developed the teacher training program, ensuring that the implementation in the field reflected the intentions and vision of the training developers. Each instructor was assigned two teacher groups from the same intervention, resulting in group sizes of 10 to 15 participating teachers.

2.2 Sampling and Randomization

The target population of our study comprises all teachers who instructed math to at least one class in grades four or five in the first year of the study. Our teacher sample is based on voluntary registration for the program. To determine our sample, we first conducted school visits in all departments to inform school heads and teachers about

[6] Beyond an initial two to three week review of the training materials, instructors had one to two weeks of preparation before each workshop with Consciente.

the trainings and the scientific study. Teachers either registered voluntarily with the representative during the school visit or later via phone. In a second step, we invited all enlisted teachers to the baseline teacher assessment, where they provided written informed consent to participate in the study.

Based on the list of participants at the teacher assessment, we selected 338 teachers to participate in the study. Following **?**, and assuming 84 teachers in the control and treatment group with an average of 18 students per teacher, 80% power, and a 5% significance level, we obtain a minimum detectable effect (MDE) of 0.125. The calculation assumes an intra-class correlation of 0.20, a between-school R^2 of 0.68, and a within-school R^2 of 0.63, calibrated using data from a prior study of teachers and students in Morazán (Büchel et al., 2022).

To avoid distortion of treatment effect estimates via spillovers (e.g., Miguel and Kremer, 2004), teachers were drawn from 338 different schools. Thus, we randomly selected only one teacher if two teachers from the same school enrolled. The resulting sample is not representative of the teacher population of our study area but mirrors standard teacher training arrangements in developing countries, where participation is often optional or weakly enforced.

Totaling 6,010 students at baseline, the student sample consists of all fourth and fifth grade students that receive math classes from the teachers in our sample. About half of the teachers only taught one of the target grades, which is why in those cases only one grade entered our sample. For the other half, one class per target grade levels four and five was part of the study.[7]

We randomly assigned teachers to either a control group or one of three experimental groups receiving: (i) math training (ii) pedagogical training or (iii) the combined approach. The random assignment was stratified by the four study departments and terciles of baseline teacher performance. Overall, this yielded twelve strata containing 28 schools on average.[8] This randomization procedure led to assignment to training of (i) 85 teachers in math, (ii) 85 teachers in pedagogy, (iii) 84 teachers in the combined approach, and finally (iv) 84 teachers receiving no training in the control group.

[7] In case a teacher instructed multiple classes on the same grade level, students of stream A, e.g. students of class 4A rather than class 4B, participated. However, this was very rarely the case.

[8] Strata leftovers were randomly assigned to treatment first within departments and finally, if still remaining as a leftover, on the national level.

2.3 Data

In all experimental groups, we conduct three measurement waves: a baseline measurement before the intervention, an endline measurement toward the end of the intervention, and a follow-up measurement one year after the intervention. This allows us to study both the immediate impact of the three interventions and whether potential effects are sustained one year after the program. Figure 3 shows the timeline of the project. The six month intervention took place in the school year 2022. The follow-up measurement was conducted at the end of 2023.

To track learning outcomes and explore potential program mechanisms, we collected an extensive set of data across teachers and students composed of: (i) teacher math and pedagogy tests, (ii) teacher classroom observation data, (iii) teacher surveys, (iii) program implementation data monitoring teacher's compliance with the program and gathering feedback on the programs, (iv) student math tests and finally (v) a short student survey.

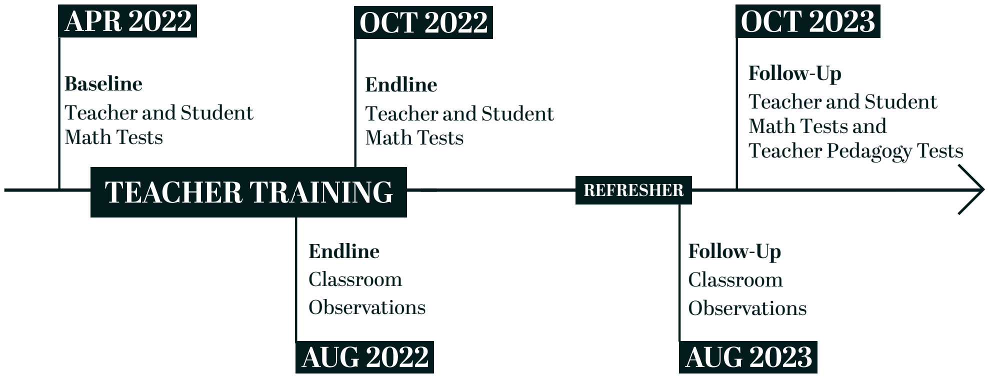


Figure 3: Timeline of the project

2.3.1 Teacher Data

As one of the two main outcomes of our study, we administered both math and pedagogy tests to evaluate teachers' knowledge of the two subjects over time. Math tests were conducted at baseline, endline and follow-up, whereas pedagogy tests were only implemented at follow-up (see Figure 3). We record a total of 338 teacher test observations at baseline, 308 at endline, and 297 at follow-up.

All **math assessments** shared a consistent structure and had a duration of 90 minutes. Each test comprised 40 grade two to grade six items, 20 of which have an overlap with the student math assessments[9]. The content of the tests was designed to be closely aligned with the composition of the curriculum, meaning that it was composed of $\sim$ 70 percent NSEA (Numeric sense and arithmetics), $\sim$ 25 percent Geometry and $\sim$ 5 percent DSP (Data, statistics and probability) questions. To avoid errors in grading the assessments were reviewed by two different people and inconsistencies resolved accordingly.[10]

The **pedagogy assessment** was divided into two sections, spanning a total of 60 minutes. The *first, main part of the test* consisted in designing a lesson plan for a 5th grade math class. Teachers were presented with a fairly detailed description of a class that has a wide age and ability range. Their job was to develop a lesson plan for a 45 minutes math class on the topic of adding and subtracting fractions. The task description included details on students previous knowledge of the topic and general guidelines for the layout of the lesson plan. Instructions advised teachers to dedicate about 50 of the 60 total minutes to this task. The *second, smaller sub-task* involved giving suggestions on how to manage and improve classroom behavior in a hypothetical setting of a 5th-grade math class, where students are restless and disruptive. The remaining 10 minutes of the test were reserved for this task. To minimize subjectivity in the grading of the tests, they were corrected twice and averaged where the outcome was a numeric, likert-scale value and re-evaluated where the outcome was binary and coders disagreed. We use principal component analysis (PCA) to create a standardized index for pedagogical knowledge based on the evaluation criteria for the pedagogical assessment. The eigenvalue distribution and corresponding scree plot shown in Figure A2 clearly meet the psychometric literature's criteria for unidimensionality. For robustness, we also present a weighted mean score of the evaluation criteria[11]. Lastly, Table A11 also presents the results by grouping evaluation criteria of the pedagogical assessments into several sub-indicators: exam completeness, general lesson implementation quality, pedagogy score, structure score, formal aspect score.

[9] Second and third grade items together have the same occurrence weight as the items from each of grades four, five, or six individually. Teacher items that intersect with student tests were picked from the general student item pool (as opposed to the student item pool of the specific wave).

[10] Items left blank were counted as incorrect.

[11] The average standardized score for part 1 of the assessment was weighed 5/6 and the average standardized score for part 2 was weighed 1/6. Criteria for part 1 were all on a scale from 1 to 3 (No, Somewhat, Yes), whereas part 2 was evaluated using the number of classroom management measures suggested and whether measures were deemed appropriate on a scale from 1 to 4. For part 2, criteria were first standardized, then averaged.

To disentangle potential mechanisms, as one of our secondary outcomes, we conduct **classroom visits**. Classrooms were monitored at endline and follow-up (see Figure 3). We record a total of 308 observations at endline and 297 at follow-up. Observations were completed using the Stallings instrument, which captures classroom snapshots at regular intervals throughout the lesson (Stallings, 1973). The seven snapshot record what the teacher is doing, what materials she or he is using, and how many students are involved in the activity. For students not involved in the teacher's activity, the recording also documents activity, material and number of students engaged. The instrument classifies as low-inference as the snapshots do not rely on the observer's personal interpretation of what is captured. As such, it yields a reliable and objective estimate of the use of the teacher's time and materials. To better link the instrument to our pedagogical interventions, we further incorporate a short set of pedagogical questions targeted at basic concepts of the pedagogical teacher training such as whether there was a lesson outline or summary and whether the teacher reviewed or gave homework.

Lastly, we collect detailed **survey and program implementation data**. To shed light on possible channels of impact, teachers were asked both at endline and follow-up about their perceived impact on students' learning, math lesson preparation, motivation and perceived competence to give a math lesson. If they taught subjects other than just math, they were additionally asked to draw comparisons to math across these aspects. Since surveys were administered immediately after teacher math assessments, the number of observations (338 at baseline, 308 at endline, and 297 at follow-up) align. To track the general intervention implementation and compliance, we keep thorough logs of teachers' attendance of on-site workshops, online meetings and self-study module completion.

2.3.2 Student Data

As the second main outcome of our study, we administer math tests to students at baseline, endline and follow-up. This allows us to evaluate whether potential learning gains in teacher knowledge translate into enhanced student performance. Exams were accompanied by a short survey that asked students for background data, their motivation and their favorite and least favorite subject as well as activity. We record a total of 6,010 student exams and surveys at baseline, 5,600 at endline and 3,050 at follow-up.

Student exams followed a structure comparable to that of the teacher assessments. Consistent with the teacher assessments, student tests included 40 questions. Due to

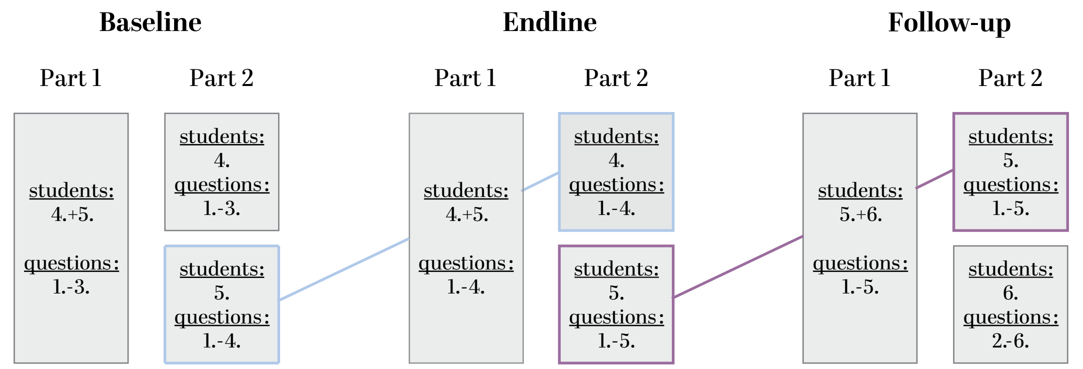


Figure 4: Assessment structure

Notes: Structure of the student assessments for each data collection wave. The numbers refer to the respective grade levels covered by the tests. Blue and purple lines illustrate assessment overlap.

the lower complexity of the items, they only lasted 60 minutes. For the estimation of the joint IRT model, items of each grade overlapped within each wave as well as over time. The first part of the test consisted of 20 items solved by all the students regardless of their grade. In a second part, tailored to each grade, students had to solve 20 more questions. As this part was grade-specific, it could be repeated in the subsequent data collection waves for the corresponding grade (see Figure 4 for a graphical representation of the exam structure). Just as with the teacher exams, the student assessments were designed to reflect the Salvadorian math curriculum in terms of topics. The tests covered between three and five grade levels.[12] In general, the distribution of items placed greater weight on higher grade levels closer to students' actual grade.[13] Lastly, in line with teacher exams, items left blank were counted as incorrect and exams were corrected by two different reviewers.

2.4 Item Response Theory

2.4.1 Linking Waves, Students, and Teachers

We leverage the overlap between tests to estimate a joint Item Response Theory (IRT) model using all teacher and student data from all waves. IRT provides a model-based alternative to percent-correct scoring by explicitly linking observed responses to an

[12] Since students at baseline were only starting their school year, they were tested on material up to the previous grade level.

[13] The two to three highest grade levels received about the same occurrence weight (depending on whether the test covered five or four grades respectively). Lower-grade items jointly appeared as often as items from a single higher grade level.

underlying latent ability θ. Unlike raw scores, which assume all items carry equal information, IRT accounts for systematic differences in item difficulty, discrimination, and (for multiple-choice items) the probability of guessing. This allows us to place teachers and students on a common scale, harmonize student scores across waves, and ensure comparability over time.

The grading framework relies on the assumption that, for each student i, the likelihood of a correct response to item j is determined exclusively by the student's latent ability θ_i and a set of item-specific parameters. We estimate the model using maximum likelihood, choosing the parameters that maximize the probability of observing the full pattern of item responses across all individuals. Under this approach, the latent trait θ_i is interpreted as the ability value that best rationalizes student i's response vector given the estimated item characteristics.

Items in our tests are dichotomously scored and can therefore be modeled using a one-, two-, or three-parameter logistic IRT specification. Let x_{ij} denote the response of student i to item j. In the three-parameter logistic (3PL) model, the probability that an individual with ability θ_i answers item j correctly is

$$P(x_{ij} = 1 \mid \theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{1}{1 + \exp[-a_j(\theta_i - b_j)]} \tag{1}$$

Here, b_j denotes the item difficulty parameter, which shifts the location of the curve and determines the level of ability at which an individual has a 50% chance of answering correctly (net of guessing). The discrimination parameter a_j governs the steepness of the item characteristic curve and thus how sharply the item differentiates individuals along the ability distribution, while c_j captures the lower asymptote, representing the probability of a correct response due solely to guessing. The one-parameter (1PL/Rasch) and two-parameter (2PL) model can be easily derived from equation 1 by constraining $a_i = 1$ and $c_i = 0$ in the case of 1PL, and relaxing the discrimination constraint ($a_i = 1$) while setting $c_i = 0$ in the case of 2PL. To improve estimation efficiency and model convergence, we employ a mixed IRT specification: we estimate discrimination and difficulty for all items, but allow the guessing parameter c_i to be freely estimated only for multiple-choice questions (about 20% of the item pool), while fixing $c_i = 0$ for open-response items. The correlation between ability estimates from this mixed model and those from alternative specifications like a full 2PL or 3PL model exceeds 99.5%, indicating that results are only minimally affected by this modeling choice.

As detailed in Chapter 2.3.2 and illustrated in Figure 4, Part 1 of the student exam provides the within-wave link, while Part 2 provides the across-wave link. Additionally, half of each teacher's test consists of items also answered by students, allowing us to jointly scale teachers and students.

Our model validation proceeds in several steps. First, we test for differential item functioning (DIF) across grades and between teachers and students. As reported in Appendix A4, very few items exhibit DIF. Second, only three items show discrimination parameters below standard thresholds.[14] Moreover, we check Cronbach's alpha for each wave and for the across-wave link to confirm internal consistency.[15] Third, excluding items with DIF or low discrimination yields IRT scores that correlate at 99.99% with our baseline estimates, further demonstrating the stability of our results. Lastly, we construct a Wright (item–person) map (see Figure A5) that plots item difficulties against the distribution of latent ability. The figure shows that item difficulties are concentrated in the upper two-thirds of the ability range, yielding high information in that region but limited precision for the lowest third of students. Estimates in the lower tail therefore depend more on extrapolation than on direct item-level variation and are thus less tightly identified.

2.4.2 TIMSS Linking

Our model incorporates a 25% overlap with the 2015 Trends in International Mathematics and Science Study (TIMSS) items (Grønmo et al., 2015), enabling us to additionally project item scores onto the TIMSS scale using IRT. This allows us to situate students' and teachers' performance in an international context and to express effect sizes on a common, externally benchmarked scale. While TIMSS primarily focuses on high-income countries, it has also been conducted in several low- and middle-income settings, such as Jordan, Morocco, and Indonesia.

To place our estimated abilities on the TIMSS scale, we adopt a common item nonequivalent groups design (Lee and Lee, 2018). In this design, two cohorts, which may differ in their underlying ability distributions, take two test forms: an old form O and a new form N. These forms include an overlapping set of anchor items. The old form is usually calibrated first, at a different point in time. The new form is then calibrated while holding the anchor item parameters fixed at their previously estimated values. In our context, the international TIMSS assessment 2015 serves as form O and our study exams serve as form N. Because the anchor items retain their

[14] Items with discrimination below 0.34 are widely viewed as weakly informative (Baker, 2017).

[15] A Cronbach's alpha of ≥ 0.8 is commonly interpreted as evidence of good internal consistency (Nunnally, 1994; Bland & Altman, 1997; DeVellis, 2016).

parameters from the TIMSS calibration, all item and ability estimates for the new form are placed directly on the TIMSS scale. Kim (2006) and Lee and Lee (2018) emphasize that fixed calibration performs well only when the latent ability distribution is allowed to adjust during estimation, rather than being fixed at the default N(0,1) distribution used by many IRT software packages. When the distribution is held fixed, the model effectively forces the new sample to match a normal distribution, causing item parameters to shift in order to accommodate any mismatch and placing new items on the wrong scale. Updating the distribution during estimation instead lets the data reveal where the new group sits relative to the anchor items, ensuring that new items are placed on the correct scale. Doing this correctly is particularly important in our setting, where students in El Salvador perform far below the TIMSS reference population.[16]

After this step, we perform a linear transformation to map the estimated IRT scores onto the actual TIMSS scale. Following the approach outlined by Patel and Sandefur (2020), we use plausible value averages from the 2015 TIMSS assessment, setting the mean as the intercept and the standard deviation as the slope parameter for the transformation.

For our mappings to be valid, the anchor items must be population invariant. In practice, this requires that the items function equivalently and measure the same construct in both populations. An intuitive illustration is a math item with substantial text: if students in our Salvadorian sample have weak reading skills, the item begins to load on a second construct, so that it no longer functions in the same way as in TIMSS. We assess population invariance by examining differential item functioning (DIF). Using the TIMSS parameter estimates, we plot the item characteristic curves for the anchor items and compare them with empirical curves for our sample. If the curves align, DIF is unlikely to be a concern. However, many of our anchor items display substantial DIF (see Figure A6), suggesting that the resulting estimates may be biased.[17] To address this, we categorize all anchor items into four groups based on the severity of DIF: substantial, moderate, very little, and none. We then estimate four alternative mappings, each using a different subset of items as linking items. One model uses the full set of TIMSS items, while the other three successively restrict the

[16] Our approach departs from Patel and Sandefur (2020), who impose a fixed ability distribution of $N(0,1)$. When we compare IRT scores under a fixed versus an updated distribution, the resulting estimates differ markedly.

[17] This is another key point of departure from Patel and Sandefur (2020). DIF plots are informative primarily for anchor items, where reference test ICCs can be compared with empirical curves under the fixed calibration. In the analysis of Patel and Sandefur (2020), only a small subset of plotted items are anchors, and these almost exclusively exhibit substantial DIF, consistent with our findings.

linking set to items in the highest-quality DIF categories. The last two categories (very little and none) contain only six and five items, which generates a familiar variance-bias trade-off: retaining fewer items reduces DIF but also yields a less stable link. The divergence across models, particularly between the unrestricted specification and the restricted ones, is substantial. When outcomes are grouped into four percentiles, the first specification agrees with the other three in only 62–78% of classifications depending on the model.[18] We select the model that retains the best three item categories for linking, thereby excluding items with substantial DIF, which constitute more than half of the TIMSS pool. This leaves 19 linking items. The resulting set strikes a balance between limiting DIF and maintaining a sufficiently large and diverse group of items that span the difficulty range of our test (see Wright Map Figure A5), reflect the distribution of TIMSS item difficulties, and remain representative of TIMSS content. As in the free model in Section 2.4.1, the linking occurs primarily in the upper half of the distribution, which again implies that scores for lower-ability students are estimated with less precision.

2.5 Balance at baseline and attrition

Table 1 confirms that randomization successfully achieved balance across experimental groups. Only one of our baseline variables for teachers and students—teacher baseline scores—shows significant differences between groups. Although teacher baseline performance was used as a stratification variable to improve balance across treatment arms, some within-stratum imbalance remains. This imbalance drives the bivariate orthogonality tests, producing differences between the control group and both T1 and T2. When teacher baseline scores are excluded from the balance checks, we no longer come close to rejecting equality across groups. To address this, we include teacher baseline scores as controls in all empirical specifications.

The average teacher is 47 years old with approximately 12 years of teaching experience, though both age and experience vary widely, ranging from 22 to 70 years and from 0 to 45 years of experience. The majority, 62% of the sample, are women and about 90% of the sample teaches other subjects in addition to math. Around 40% of teachers instruct math in both 4th and 5th grade, and the average class size is 16 students. On average, teacher answered about 61% of the baseline math items correctly (for a more detailed analysis of teacher and student baseline performance see Section 3.1).

[18] The discrepancy is even larger when the latent ability distribution is fixed at $N(0, 1)$ rather than updated during estimation, with agreement falling to 35–53%.

Student performance at baseline was relatively low, with an average score of 29%. Math is the favorite subject for about 40% of students, while another 40% rank it among their least favorite. The sample is evenly split between male and female students, and the average student is about 11 years old. The typical student lives in a household with 5 members, including the child itself. The socioeconomic status is measured as the average of binary indicators for household assets, including electricity, a TV, a computer, a fridge, internet access, a car, a washing machine, and books. On this index, the average household scores approximately 70%.

Table 1: Balance at Baseline

	Control	T1: Math	T2: Pedagogy	T3: Combined	P-value	N
Teacher variables						
Baseline score	0.59	0.61	0.63	0.61	0.04	338
	(0.20)	(0.22)	(0.22)	(0.21)		
Sex	0.64	0.62	0.59	0.64	0.86	338
	(0.48)	(0.49)	(0.50)	(0.48)		
Age	45.60	47.90	46.86	46.14	0.42	336
	(10.36)	(10.82)	(9.63)	(10.12)		
Class variables						
Class size	12.46	13.19	11.82	12.69	0.66	334
	(7.35)	(7.22)	(6.71)	(6.51)		
Student variables						
Baseline score	0.29	0.29	0.28	0.30	0.30	6010
	(0.16)	(0.16)	(0.15)	(0.16)		
Favorite subject math	0.43	0.41	0.42	0.42	0.88	5399
	(0.49)	(0.49)	(0.49)	(0.49)		
Female	0.49	0.50	0.51	0.49	0.42	7122
	(0.50)	(0.50)	(0.50)	(0.50)		
Age	10.62	10.67	10.71	10.68	0.91	6722
	(1.17)	(1.25)	(1.14)	(1.23)		
SES index	0.14	0.02	-0.01	0.05	0.24	6573
	(1.40)	(1.41)	(1.50)	(1.51)		
Household members	5.33	5.33	5.26	5.25	0.48	7068
	(2.12)	(2.29)	(2.16)	(2.14)		
Orthogonality Tests						
Bivariate: Control vs Math					0.166	
Bivariate: Control vs Pedagogy					0.059	
Bivariate: Mentoring vs Combined					0.674	

Standard errors in parentheses clustered at the teacher level. Individual p-values are from F-tests that treatment status does not predict each individual outcome variable, calculated controlling for age, gender, and stratum fixed effects, except for Age (does not include age fixed effects) and Female (does not include gender fixed effects). Bivariate orthogonality tests examine whether baseline variables predict treatment assignment in pairwise comparisons between each treatment arm and the control group. Orthogonality tests are computed using randomization inference (1000 replications) which mitigates concerns about over-rejecting when many balance variables are included (see XX) Orthogonality tests include stratum fixed effects. All standard errors are clustered at the school level.

Table 2 presents attrition rates for teacher assessments, classroom observations, and student assessments across the data collection waves. Endline attrition was low and balanced across treatment arms, at 9 percent for teachers (tests and classroom observations) and 18 percent for students. Attrition at follow-up remained low for teacher tests (12 percent) but was considerably higher for student assessments (59 percent). This reflects the frequent changes in teacher timetables experienced in El Salvador, where teachers often face reassignments or shifts in responsibilities. Specifically, we lost 123 entire teacher clusters and approximately 2,700 students due to teachers being assigned to new classes, no longer teaching math, or being promoted to school director positions. These changes also impacted attrition in classroom observations, though to a lesser extent. For classroom observations, we were able to maintain slightly lower attrition rates because data collection did not require teachers to instruct the same class as in the baseline and allowed observations in grades 3 through 9 when no fourth- or fifth-grade class was available. Despite these challenges, we find no strong evidence of differential attrition by treatment status that would threaten the validity of our results, as detailed in Table A4.[19]

We consider whether the loss of statistical power stemming from attrition at follow-up may constrain inference. We therefore revisit the original power calculations reported in 2.2 and reestimate minimum detectable effects (MDEs), adjusting the number and size of clusters to the realized follow-up sample while holding all other parameters fixed. These revised estimates indicate that, despite substantial attrition, the study remains sufficiently powered to detect effects of 0.153σ–0.162σ with 80 percent probability at the 5 percent significance level.

Table 2: Attrition per Data Collection Type and Wave

	Attrition Rates		*Number of Observations*	
	Endline	Follow-up	Endline	Follow-up
Teacher tests	0.09	0.12	309	297
Classroom observations	0.09	0.32	309	231
Student tests	0.18	0.59	4950	2441

[19] Attrition is slightly lower (by 10–15%) in the combined treatment arm (T3) for student tests. We find no plausible mechanism linking treatment assignment to this pattern and this does not affect the interpretation of the results in Section3.

3 Results

3.1 Descriptive Results

Panel (a) in Figure 5 presents the results of the baseline math assessments for teachers. The average teacher is able to correctly answer 61% of grade two to six questions. Teachers' highest performance was observed in basic numeric operations and more mechanical tasks like subtraction with two digits and fraction addition, where they achieved around 70%–90% correct responses. Similarly, they performed well in basic geometric tasks like converting units or reading time (∼90%). However, their performance declined considerably in more complex concepts requiring higher-order reasoning. For instance, only 40% correctly calculate a simple percentage, less than a third understands the order of operations (30%), or how to determine the volume of a cube (31%) and only about one out of six can solve a simple inverse proportionality problem (17%). Panel (a) from Figure 6 further illustrates this pattern, showing a sharp decline in correct answers as question difficulty increases beyond grade three. On average, teachers solved fewer than half of the grade five questions correctly (49%) and only about a third of grade six questions (38%). The observed knowledge gaps suggest that the majority of teachers lack a firm command of the primary curriculum, leaving many in the difficult position of having to teach material they have not fully mastered themselves.

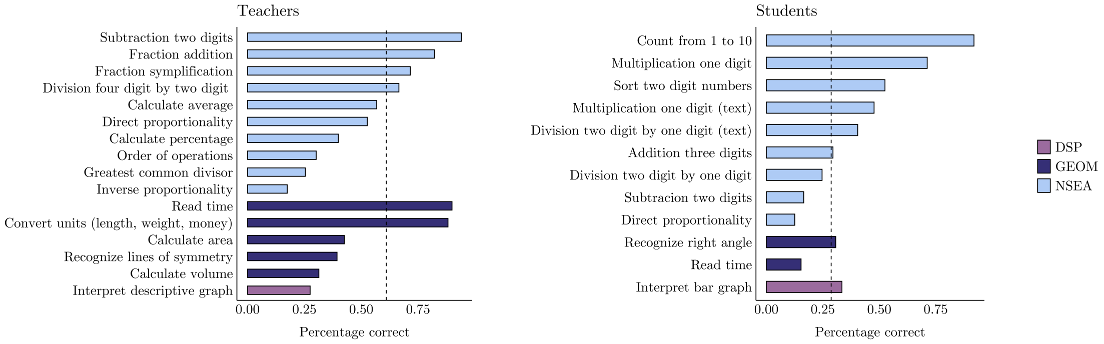


Figure 5: Average percentage per item examples

Notes: The two figures compare mean performance on selected baseline assessment items across all treatment groups. The figures present teacher results on the right and student results on the left. On average, teachers scored 61% correct and students 29%.

Panel (b) in Figure 5 paints a similarly concerning picture for 4th and 5th grade students, with even lower overall performance. On average, students correctly answered only 29% of the math baseline questions, which assessed grade one to three content aligned with the curriculum. As with teachers, students performed better on the simplest tasks—such as counting from 1 to 10 or solving basic one-digit arithmetic—where accuracy ranged between 50% and 90%. However, performance dropped sharply for slightly more advanced problems. For example, only one out of four students knows how to solve an addition problem involving a three-digit number, one in five correctly divides a two digit number by a one digit number, and just one out of six students knows how to subtract two digit numbers from each other. Fifth graders slightly outperform fourth graders (see panel (a) in Figure 6), suggesting that an additional year of schooling helped reinforce at least some foundational skills. However, the decline in performance as item difficulty increases is even steeper for students than for teachers, pointing to a clear deficiency in fundamental math competencies that are critical for progression in higher grades.

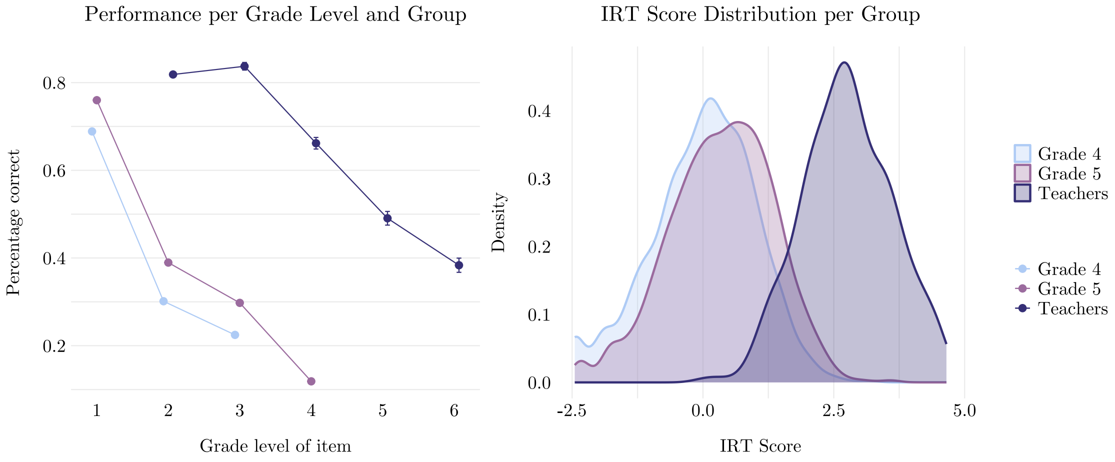


Figure 6: Average baseline performance per group

Notes: The left figure shows percent correct per grade level of the items for teachers and students. The right figure illustrates the distribution of baseline irt scores for each teachers and students. IRT scores are standardized to the 4th grade baseline distribution.

By concurrently calibrating the IRT model with both teacher and student responses, we are able to directly compare students and teachers latent traits at baseline. Panel (b) from Figure 6 illustrates the distribution of these traits across groups. While teachers, as expected, generally outperform students, there remains a sizable degree of overlap between student and teacher distributions. Approximately 57 percent of students outperform the worst-performing teacher, and 0.2 percent outperform the median teacher. This overlap, however, is primarily driven by variation across clusters, as only 72 out of roughly 6,000 students outperform their own teacher. This suggests that student ability is largely constrained by teacher ability—a student can only achieve as much as their teacher enables them to.

To place Salvadorian students' and teachers' abilities in an international context, we map their performance onto the TIMSS scale by leveraging the overlap between the TIMSS exam and our assessments (see section 2.4). Figure 7 presents this mapping, reinforcing our earlier descriptive findings. The TIMSS test in 2015 was conducted with fourth graders from 49 countries, most, but not all, of which are high-income countries. Since TIMSS is given later in the academic year than our assessments, the true correspondence between our fourth- and fifth-grade scores and the TIMSS benchmark plausibly lies between the two grades. El Salvador's fourth and fifth graders rank near the bottom of the international distribution, positioned just above Kuwait, the lowest-performing country in the sample. Among the nine middle-income countries that took part in the TIMSS assessment, El Salvador scored lowest. Even more strikingly, El Salvador's teachers only very narrowly outperform Singapore's fourth graders—the highest-scoring students in the assessment. That teachers struggle to surpass 10-year-olds from top-performing systems is yet again a staggering reflection of the profound knowledge gaps within El Salvador's education system.

A complementary benchmark comes from the harmonized learning outcomes database of Altinok et al. (2018), which places national assessments on the TIMSS scale using simple linear scaling. The approach exploits countries where students participate in both regional and international assessments, generating the necessary link for the transformation. Unlike Patel and Sandefur (2020), our estimates align quite closely with the harmonized learning outcomes database, which reports an average score of 364. Our mapped scores of 365 for grade 4 and 396 for grade 5 lie close to this reference value. The comparison is also consistent when examining proficiency cutoffs. Using the TIMSS Low International Benchmark of 400 points, Altinok et al. (2018) report that 51 percent of students reach this threshold, compared with 35 and 49 percent of fourth and fifth graders from our sample, respectively.[20] For the advanced benchmark of 625 points, the database reports 0.46 percent of students reaching this level, versus 0.17 and 0.36 percent in our data.

3.2 Experimental Results

To assess the causal effect of the three treatments on teacher and student outcomes for each post-treatment $wave \in \{endline, follow-up\}$, we use

[20] For context, among participating countries and territories in TIMSS 2019, 92 percent of fourth-grade students reached the Low International Benchmark in mathematics (Mullis et al., 2020)

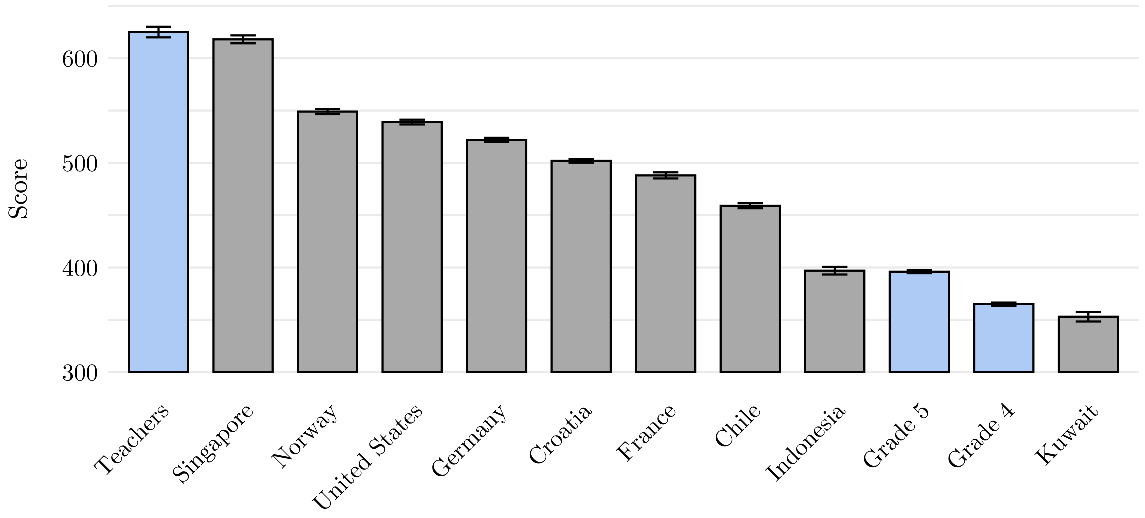


Figure 7: El Salvador in International Comparison: TIMSS

Notes: The blue bars represent teachers' and students' test scores mapping from El Salvador using our own data. The TIMSS test in 2015 was conducted with fourth graders from 49 countries, most of which are high-income countries. Among these countries, Singapore achieved the highest scores, while Kuwait scored the lowest. The figure highlights additional countries that provide meaningful points of comparison.

$$Y_{ic}^{wave} = \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3 + \delta Y_{ic}^{baseline} + \mu_s + \epsilon_{iv} \tag{2}$$

where Y_{ic}^{wave} are endline or follow-up outcomes for student i of teacher c; T_1, T_2 and T_3 are treatment indicators for treatment 1 (math training), treatment 2 (pedagogy training) and treatment 3 (combined training); $Y_{ic}^{baseline}$ is the baseline math test score; and μ_s are strata fixed effects. Standard errors are clustered at the teacher level. For teacher outcomes, we extend Equation 2 by adding individual-level controls for sex and age.

Table 3 presents the estimated impacts on student learning outcomes immediately after the program and one year later. None of the three treatment groups exhibit statistically significant differences in math scores relative to the control group at either point in time. The small point estimates and standard errors suggest that the true effects are close to zero.[21]

[21] A notable exception is the marginally significant negative coefficient for the combined treatment on the TIMSS scale at endline (-8.028, $p < 0.1$), which becomes statistically insignificant when controls are added (-6.922, $p = 0.16$). This finding is somewhat surprising given that neither the math-only nor pedagogy-only treatments show negative effects, and the combined treatment represents a blend of both approaches. Three explanations are conceivable: (1) negative complementarities between content and pedagogical training, though we cannot identify a plausible mechanism for this;

When examined separately by grade, the immediate post-intervention effects show slightly positive point estimates for grade 4 and slightly negative ones for grade 5, though these estimates remain statistically insignificant. By the one-year follow-up, the grade-specific point estimates converge back toward zero, reinforcing the conclusion that the intervention did not meaningfully affect students' math achievement.

Table 3: ITT estimates: Student math scores

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Without Controls								
Math score (sd)	0.030	0.002	-0.101	4950	-0.000	-0.011	-0.035	2441
	(0.067)	(0.076)	(0.065)		(0.083)	(0.087)	(0.102)	
Math score (IRT)	0.010	-0.042	-0.096	4950	0.014	-0.043	-0.029	2441
	(0.062)	(0.072)	(0.061)		(0.078)	(0.086)	(0.095)	
Math score (TIMSS)	1.004	-2.998	-8.028*	4950	0.959	-3.457	-2.641	2441
	(4.893)	(5.728)	(4.877)		(6.466)	(7.014)	(7.887)	
With Controls								
Math score (sd)	0.043	0.013	-0.082	4366	0.033	0.021	0.008	2161
	(0.064)	(0.076)	(0.065)		(0.084)	(0.087)	(0.102)	
Math score (IRT)	0.021	-0.033	-0.082	4366	0.053	-0.012	0.007	2161
	(0.059)	(0.070)	(0.061)		(0.078)	(0.085)	(0.093)	
Math score (TIMSS)	2.054	-2.364	-6.922	4366	4.046	-0.881	0.446	2161
	(4.679)	(5.618)	(4.885)		(6.446)	(6.919)	(7.732)	

All the dependent variables are standardized except for the TIMSS scores. Controls include strata fixed effects, the outcome variable at baseline and demographic controls. Standard errors are clustered at the teacher level. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

4 Discussion

Our theory of change posits four critical intermediary steps necessary for improving student test scores. *In a very first step*, teachers must attend the training and the program must be implemented with fidelity. *Second*, teachers must achieve meaningful

(2) the 50% reduction in time devoted to each component relative to the standalone treatments may have introduced pedagogical techniques without sufficient depth, creating confusion rather than learning gains; or (3) a chance finding. Given the lack of persistence at follow-up, the absence of significance in the controlled specification, and the small magnitude (approximately 0.1 standard deviations), we consider explanation (3) most likely.

learning gains to have the requisite knowledge and skills to transmit to their students. This implies that improvements in teacher knowledge, be it in content or pedagogical knowledge, must not only be detectable but also substantial enough to influence student learning. *Third*, particularly in the case of pedagogical competencies, but also for content knowledge through explanatory quality, feedback and pacing, these newly acquired skills must be successfully integrated into classroom instruction. *Fourth*, even if teachers successfully enhance knowledge and their practice, student learning outcomes will only improve if students can effectively absorb, process, and retain the newly introduced content. The extent to which these conditions are met determines whether the intervention leads to measurable gains in student achievement.

4.1 Compliance, Implementation, and Power

We first show that our intervention was implemented with high fidelity and avoided the early compliance and delivery failures that derailed comparable government programs and blocked the first and necessary step of improving teacher knowledge and classroom practice (Loyalka et al., 2019; Schaffner et al., 2025).

Table 4 summarizes program participation. Over all treatment groups, 81% of teachers attended at least one of the seven in-person workshops with the average teacher completing 4.75 out of the seven workshops (including non-participants). Participation rates for online self-study modules and required homework submissions were similarly high (see Table 4). Consistent with this pattern, teachers in every treatment arm reported significantly greater perceived workload, rising by roughly 0.5 standard deviations relative to the control group ($p < 0.01$; Table A5). Participation is slightly lower in the combined and pedagogy treatments relative to the math group, as reflected in the number of completed workshops ($p = 0.11$ and $p < 0.01$). This aligns with field reports suggesting that teachers appeared less familiar with the concept of pedagogy and were primarily motivated to participate by the availability of math training. Participation remained stable over time, with no clear patterns or variations across all treatments (see Figure A3).

Loyalka et al. (2019), the first of two identified studies that assess classical government training programs in primary schools but find null effects for students and teachers, attribute the failure of their program to content that was overly theoretical and delivered in a rote, non-interactive manner. In designing our intervention, we explicitly sought to avoid such weaknesses: as described in Section 2.1, the curriculum adhered to rigorous pedagogical standards, was tailored to teachers' knowledge

Table 4: Compliance by Treatment

Treatment	Attended one or more (%)	Completed workshops	Completed online activities	Completed homework sets
T1: Math	0.85	5.36	5.06	5.59
T2: Pedagogy	0.82	4.69	4.46	5.08
T3: Combined	0.77	4.18	4.15	4.35

The total number of activities for each block is 7.

levels, and emphasized interactive learning. Twelve semi-structured interviews with a randomly selected subset of teachers across treatment arms indicate that these features were salient. Teachers consistently valued the training, highlighted the quality of instruction, and conveyed a clear sense that the sessions differed positively from traditional government trainings (see Table A7 for exemplary quotes from the interviews). Many also provided concrete examples of how the interactive methods altered their classroom practice. When asked how the training could be improved, teachers generally offered no suggestions and replied that "everything was great".[22]

Schaffner et al. (2025), the second study on classical government training, identify low instructor motivation, inadequate preparation and support for instructors and weak subject knowledge of instructors as key barriers to generating meaningful teacher and student impacts. To address these concerns, our intervention invested heavily in instructor selection and support. Facilitators were young teachers hired by the implementing NGO, each with at least a bachelor's degree in mathematics or educational science (3-6 years oof university instruction) and relevant classroom experience. Of seventeen pre-screened candidates, twelve were ultimately hired following subject-matter testing and 1.5 weeks of initial training. Upon hiring, they received two additional weeks of general preparation as well as block-specific training delivered by the respective Salvadorian expert ahead of each instructional block. Their schedule also allowed roughly two weeks of preparation time per block. During the intervention, facilitators received weekly in-person and online coaching from Salvadorian and Swiss experts and two monitoring visits to ensure quality. According to NGO reports, all twelve facilitators demonstrated strong motivation and performance throughout.

[22] A word cloud of responses to the improvement question shows that "everything" was the most frequently mentioned term, and it appeared only alongside positive descriptors such as "great" and "amazing" (Figure A10).

4.2 Teachers' Learning Gains

Unlike Schaffner et al. (2025), who posit that, aside from instructor-related issues, their training did not target teachers' actual knowledge levels, we find that our intervention does so and leads to significant gains in teacher knowledge. Table 5 shows effect sizes for teachers' content and pedagogical knowledge. Immediately after the intervention, teachers in the math treatment group improved their scores by 0.14σ ($p < 0.05$). These effects increased to 0.29σ ($p < 0.01$) one year later, suggesting that teachers continued consolidating their knowledge over time. This effect is sizeable, corresponding to a 24-point shift on the TIMSS scale. It is comparable to the gap between Salvadorian fourth and fifth graders and to the difference between the fourth-grade country averages of Chile and France. As expected, there were no significant impacts on math scores for teachers who received only the pedagogy training. While a small negative coefficient appears immediately after the program, it is insignificant and shifts to a positive, yet still insignificant, value in the follow-up measurement. For teachers in the combined training group, we observe an initial improvement of 0.11σ relative to the control group, although the effect falls just short of statistical significance ($p = 0.14$). Similar to the math-only group, the effect increases to 0.15σ in the follow-up ($p < 0.1$), indicating that these teachers also strengthened their math skills over time.

Table 5: ITT estimates: Teacher scores

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Teachers: Math								
Math score (sd)	0.142** (0.072)	-0.095 (0.074)	0.108 (0.073)	308	0.291*** (0.077)	0.049 (0.079)	0.146* (0.078)	297
Math score (IRT)	0.139* (0.074)	-0.144* (0.076)	0.088 (0.075)	308	0.291*** (0.082)	0.063 (0.085)	0.136 (0.084)	297
Math score (TIMSS)	12.523** (6.240)	-10.395 (6.426)	7.357 (6.310)	308	24.111*** (6.516)	5.299 (6.701)	11.522* (6.658)	297
Teachers: Pedagogy								
Pedagogy Score (PCA)	-	-	-	-	0.491*** (0.177)	0.573*** (0.182)	0.466** (0.180)	295
Pedagogy Score (Weighted Mean)	-	-	-	-	0.453*** (0.172)	0.550*** (0.176)	0.428** (0.175)	297

All the dependent variables are standardized. Controls include strata fixed effects and the outcome variable at baseline. Results for teachers include demographic controls. Standard errors are clustered at the teacher level. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

We find substantial and statistically significant long-term effects on pedagogical knowledge, as measured by the PCA index, across all three treatment groups. The largest effect, 0.57σ ($p < 0.01$), is observed among teachers who received the pedagogical training. The effects for the math and combined treatments are only slightly smaller, at 0.49σ ($p < 0.01$) and 0.47σ ($p < 0.05$), respectively. This indicates that all three interventions led to lasting improvements in pedagogical knowledge even one year after the program ended. The differences in effect sizes between the pedagogical-training group and the other two treatment groups are not statistically significant ($p = 0.83$). These findings remain robust when using a weighted mean index instead of the PCA-based measure.

A remaining question is whether the magnitude of the observed gains in teachers' math content knowledge, specifically, was sufficient to translate into student learning. While effects for teachers in the math training reach 0.26σ for specific math domains (see Table A12) and doubles at follow-up, it remains plausible that the overall improvement was insufficient for teachers to effectively transfer this knowledge to their students. This concern aligns with quasi-experimental evidence from Peru (Metzler and Woessmann, 2012), several African countries (Bietenbeck et al., 2018; Bold et al., 2019), Pakistan (Bau and Das, 2020), and El Salvador (Brunetti et al., 2024), which indicates that a 1σ increase in teacher content knowledge is associated with only a 0.1σ improvement in student learning outcomes. Our own data corroborate this correlational relationship suggesting that a 1σ increase in teacher content knowledge is associated with a 0.08σ higher student performance (see Table A14).

However, the math treatment extended beyond improving teachers' content knowledge; it also led to a substantial increase of approximately 0.5σ in their pedagogical knowledge. This effect likely stems from the training's interactive and pedagogically structured design. When mathematical concepts are taught to teachers using pedagogically sound methods, it is reasonable to expect that teachers will replicate these approaches in their own classrooms—a mechanism that is consistent with evidence from our semi-structured interviews. This additional pedagogical component of the math treatment suggests that, despite arguably moderate effects on content knowledge, the intervention's impact on student learning may not be strictly constrained by the teacher-student knowledge transmission patterns established in section 3.2.

4.3 Classroom Implementation

Given these gains in teachers' pedagogical content knowledge, we next examine whether they translate into changes in classroom instruction. We find that teachers integrate newly acquired pedagogical skills into classroom instruction in the short run. Specifically, we observe a shift in classroom practices at endline for the treatment groups that received a pedagogical component in their training (see Table 6). Teachers in these treatment groups exhibit a greater reliance on interactive materials, including shared resources, learning aids, and information and communication technology, using them approximately 5 percentage points more frequently than their control group counterparts ($p < 0.1$). At the same time, their reliance on conventional materials, such as notebooks, textbooks, and blackboards, declines by 8 percentage points ($p < 0.05$).

Teachers in both pedagogical treatment groups also perform significantly better in our constructed pedagogical practices index, scoring approximately 0.35σ higher than their peers in the control group ($p < 0.05$). This index measures instructional quality through six key dimensions, including lesson structuring, assignment and review of homework, lesson summarization, teacher mobility around the classroom, and use of examples. Additionally, teachers in these groups are 18.4 and 12.8 percentage points more likely ($p < 0.01$ and $p < 0.05$, respectively) to implement interactive activities, representing a 200% and 140% increase relative to the control group mean. Improved instructional practices are also associated with enhanced student engagement, as evidenced by a 9 percentage point reduction in classroom distractions, amounting to a 20% improvement over the control group mean.

However, these gains prove to be short-lived. Effects dissipate within a year, indicating that training alone may be insufficient to induce long-term behavioral change. Anecdotal evidence from developing countries frequently highlights the persistence of rigid teaching practices, where teachers adhere closely to prescribed curricula (**?**). As a result, they may display resistance to pedagogical shifts, even when given the necessary tools to implement them. This challenge may be particularly relevant in the case of El Salvador, where the government introduced a standardized mathematics textbook, ESMATE, in 2018. Designed to align with the national curriculum, ESMATE offers scripted lesson plans that teachers are expected to follow. While experimental evaluations have demonstrated positive learning outcomes associated with the textbook (Maruyama and Kurosaki, 2024), its structured nature may limit teachers' flexibility in tailoring lessons to their students' needs.

Empirical evidence from our teacher survey supports this concern. As illustrated in Figure 8, 90% of teachers report always or very often preparing their math lessons

Table 6: Classroom observations: Teachers

	Endline			*Follow-up*		
	T1: Math	T2: Ped	T3: Both	T1: Math	T2: Ped	T3: Both
Teacher use of time						
Active (%)	-0.013	-0.002	0.014	-0.014	-0.052	0.027
	(0.036)	(0.036)	(0.036)	(0.039)	(0.039)	(0.039)
Passive (%)	0.006	0.035	-0.017	0.028	0.050	0.001
	(0.035)	(0.036)	(0.036)	(0.033)	(0.033)	(0.033)
Management (%)	0.014	-0.018	0.010	0.035	0.044	0.005
	(0.024)	(0.024)	(0.025)	(0.028)	(0.028)	(0.028)
Material use						
Good Material (%)	0.039*	0.051**	0.047*	0.034	0.022	0.030
	(0.023)	(0.024)	(0.024)	(0.025)	(0.025)	(0.025)
Bad/Non-interactive Material (%)	-0.057	-0.078**	-0.082**	-0.061	-0.038	-0.028
	(0.038)	(0.038)	(0.039)	(0.040)	(0.041)	(0.041)
Pedagogy						
Pedagogical score (sd)	0.140	0.335**	0.341**	-0.170	0.111	-0.118
	(0.168)	(0.169)	(0.170)	(0.153)	(0.154)	(0.154)
Activity variation (sd)	0.200	-0.025	0.059	0.173	-0.070	0.072
	(0.141)	(0.142)	(0.144)	(0.161)	(0.162)	(0.162)
Conducts interactive activity	0.010	0.184***	0.128**	0.107**	0.123**	0.050
	(0.058)	(0.058)	(0.059)	(0.051)	(0.051)	(0.051)
One or more students distracted (%)	-0.010	-0.090*	-0.048	-0.007	-0.061	-0.023
	(0.046)	(0.047)	(0.047)	(0.047)	(0.047)	(0.047)

Sample sizes are n = 307 for the estimation of endline effects and n = 230 for follow-up effects. All the dependent variables are either standardized (sd), binary or indicated as a percentage of the lesson time (%). Good material constitutes shared material, learning aids or ICT. Bad/Non-interactive material consitutes textbook, notebook and blackboard. The pedagocical score is a summed scored of the following standardized variables: Does lesson outline, does homework review, uses examples, moves in classroom, gives homework, does lesson summary. Controls include strata fixed effects and demographic teacher variables.

by consulting the ESMATE lesson plan, and all teachers in the sample rely on the textbook as their main instructional guide. Notably, the intervention did not alter teachers' reliance on ESMATE, as there is no evidence of differential usage across experimental groups. These findings suggest that even when professional development programs enhance teachers' pedagogical knowledge, institutional structures—such as mandated instructional materials—may constrain the translation of these improvements into more adaptive and responsive teaching practices.

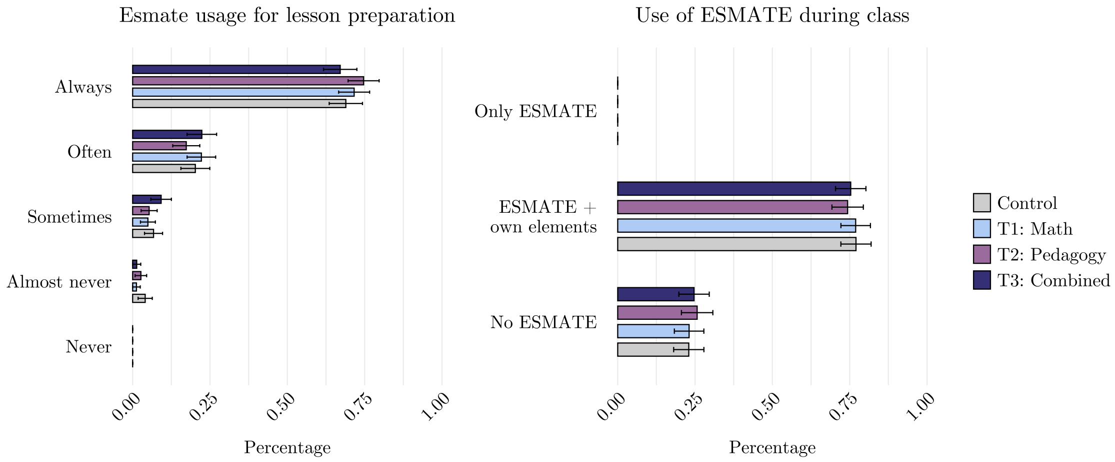


Figure 8: Use of ESMATE during class

Notes: The left figure illustrates how frequently teachers use ESMATE for lesson preparation, while the right figure shows the extent of ESMATE usage during classroom instruction. Both measures are based on self-reported data provided by teachers during the endline data collection. There is no differential use of ESMATE across treatment groups (p-value of a joint F-Test that each treatment dummy coefficient is equal to 0 = 0.98).

Considering our findings from Figure 8 and Table 6, we interpret these results as presumptive evidence that teachers revert to their established routines once external reinforcement is removed. This suggests that pedagogical training might be more successful if it is paired with long-term support such as pedagogical coaching that extends beyond training to consolidate initial training gains and prevent regression to pre-intervention teaching practices. By the same logic, training multiple teachers within a school, rather than only one, could create a more supportive environment for implementing new practices across classrooms.

4.4 Student Deficit

As discussed in Sections 4.2 and 4.3, teacher effect sizes and classroom implementation can pose challenges to the successful delivery of teacher training interventions. However, even when these factors are not the primary constraints on student learning, as suggested by the endline results, students may still face significant obstacles in learning, understanding and retaining the knowledge passed on to them. This can happen if there is a misalignment between the curriculum and their actual competencies.

Section 3.1 illustrates this point. Many students perform well below grade-level expectations: the average fourth grader answers 22 percent of grade-level questions correctly, and the average fifth grader answers 12 percent. Consistent with these patterns, IRT scores place fourth graders at roughly a first-grade level and fifth graders between first- and second-grade levels.

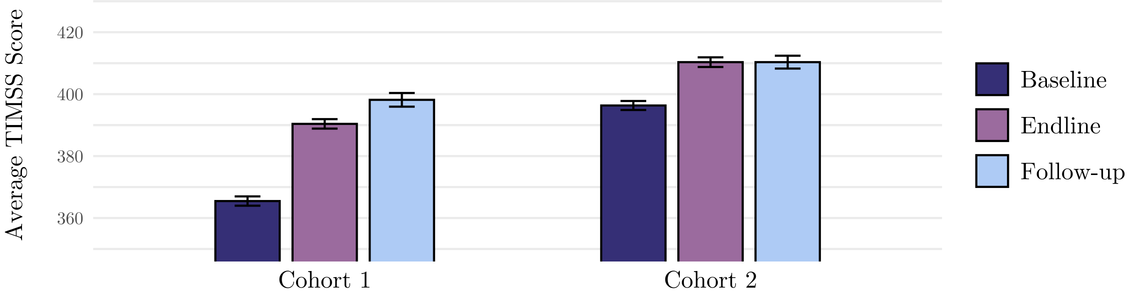


Figure 9: Student learning per grade

Notes: The first cohort corresponds to students who were in grade 4 at baseline and endline and grade 5 at follow-up. Similarly, the second cohort represents students who were in grade 5 at baseline and endline and grade 6 at follow-up. Scores are shown on the estimated TIMSS scale.

At the same time, quasi-experimental evidence suggests that learning trajectories in primary education in developing contexts often follow a concave curve (Patel and Sandefur, 2020; Banerjee et al., 2016). As children progress through grades, their learning gains diminish each additional year, as foundational skills acquired in the early years remain underdeveloped. Once students fall too far behind, they may derive little to no benefit from grade-level instruction (see (Muralidharan et al., 2019)). Using scores on the TIMSS scale, which are independent of test difficulty, Figure 9 confirms this pattern in our data: average value added per year of schooling declines as cohorts advance through grades within our study period. Students in grade 4 exhibit significantly larger learning gains over the school year than those in grade 5. For cohort 2, learning gains appear to stagnate in grade 6, indicating that the marginal value added of this grade may be minimal.

Taken together, these patterns suggest that many students may be too far behind the curriculum to benefit from improved grade-level instruction. Evidence from our endline supports this hypothesis. Figure 10 presents polynomial estimates of treatment effects as a function of students' distance from grade-level proficiency at baseline. Distance is defined relative to the average item difficulty for the relevant grade, with zero indicating performance at grade level; a distance of approximately 0.4 corresponds to one grade level on average.[23] The figure shows that, immediately after the

[23] For the transition from grade 1 to grade 2, the implied distance is substantially larger (approximately 1). This may reflect a steeper increase in curricular difficulty between these grades or greater

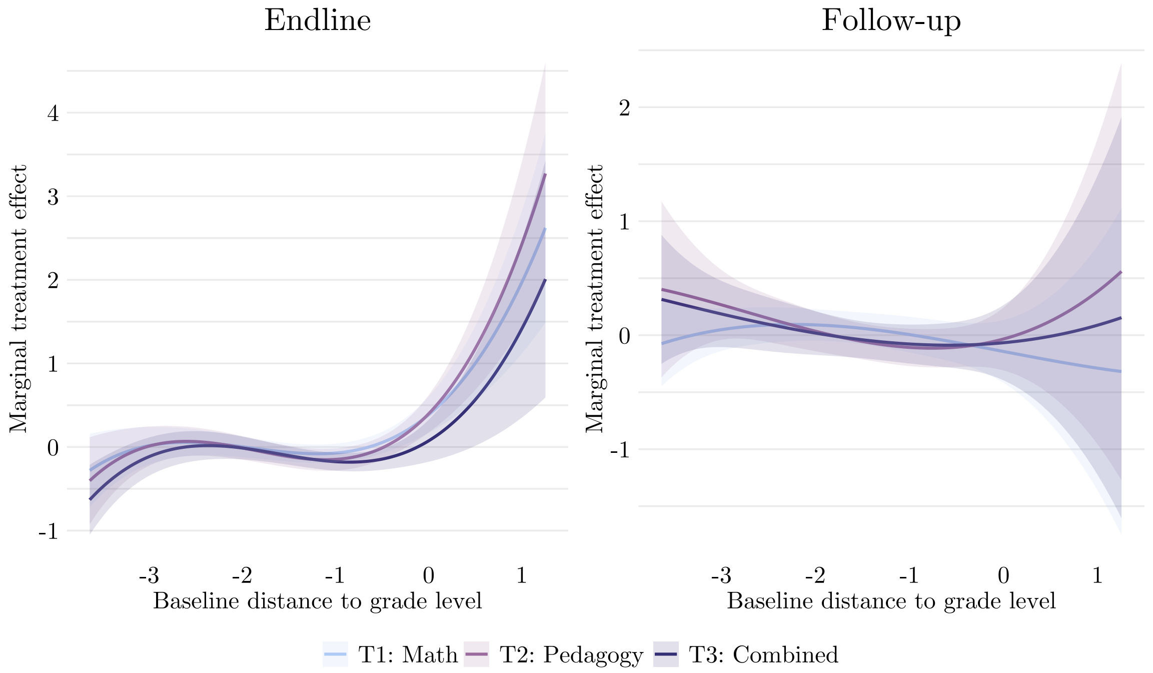


Figure 10: Endline effect heterogeneity over baseline student performance

Notes: Baseline distance to grade-level proficiency is measured relative to the average difficulty of grade-level test items. The figure plots marginal treatment effects estimated from regressions interacting treatment indicators with a third-order polynomial in students' baseline distance; shaded areas denote 90% confidence intervals.

intervention, students who were at or above grade level at baseline benefited across all three treatment arms. Hereby, a distance of 0.4 represents about one grade level on average, meaning that children who are at -0.4 are about one grade level below their actual level.

In settings characterized by large learning gaps, these findings suggest that teacher training programs focused on improving grade-level instruction may face inherent limitations. By contrast, early interventions that strengthen foundational skills may be critical for sustaining learning progress over time and preventing the plateauing of learning trajectories in later grades. This aligns with prior studies, which have predominantly identified positive effects of teacher training in early primary grades, particularly in early reading programs or foundational skill interventions (Kerwin and Thornton, 2021; Piper et al., 2018).

measurement noise, as grade 1 difficulty is estimated from a smaller set of test items.

5 Conclusion

Student learning levels remain critically low in many developing countries—a challenge in which teacher quality emerges as a pivotal factor. Using a large-scale randomized controlled trial in El Salvador, we evaluate three intensive teacher training programs targeting subject knowledge, pedagogical skills, or both. The programs succeed where conventional in-service and government-led trainings have fallen short, producing sizable improvements in teacher knowledge and short-term classroom implementation (Loyalka et al., 2019; Schaffner et al., 2025). Yet these improvements do not lead to measurable gains in student learning.

Our results identify two interrelated constraints on program effectiveness, even under successful implementation. First, structural factors, such as mandated curricular materials and pacing, may constrain teachers' ability or willingness to incorporate newly acquired pedagogical strategies, potentially limiting the program's impact on daily teaching practices. Understanding and adapting to such constraints is critical for designing future policies that balance curricular structure with instructional flexibility to maximize learning outcomes.

Second, effective teaching is limited by a substantial misalignment between the primary curriculum and students' underlying skill levels. Fourth- and fifth-grade students correctly answer just over 20 percent and 10 percent of items from the preceding grade, indicating large cumulative learning gaps. These deficits are also evident in international comparisons, where Salvadorian students rank near the bottom of the TIMSS distribution. In contexts where students lag fare below grade-level expectations, even well-trained teachers may struggle to generate sustained improvements if they are expected to teach at their grade level.

The findings of this study highlight several important policy implications for improving the effectiveness of teacher training programs. *Firstly*, training programs must be well-aligned with the specific structural constraints of each educational setting to ensure relevance and feasibility. *Second*, sustaining initial gains and preventing reversion to prior teaching practices may require pairing pedagogical training with longer-term support, such as ongoing instructional coaching. This interpretation is consistent with evidence showing that structured pedagogy interventions rank among the most effective education interventions tested (see, for example, Angrist et al. (2025)). Relatedly, training only one teacher per school may not be sufficient to instigate widespread change; training multiple teachers within the same school could foster peer interaction and facilitate the diffusion of new teaching practices across classrooms. *Finally*, and perhaps most importantly, it is crucial to focus teacher training or other educational interventions on mitigating early learning deficits. This is when learning gaps typically first emerge and if left unaddressed, can persistently affect a child's educational trajectory as well as the effectiveness of teacher training programs.

3 References

Alan, S. and Mumcu, I. (2024). Nurturing childhood curiosity to enhance learning: Evidence from a randomized pedagogical intervention. *American Economic Review*, 114(4):1173–1210.

Albornoz, F., Anauati, M. V., Furman, M., Luzuriaga, M., Podesta, M. E., and Taylor, I. (2020). Training to teach science: experimental evidence from argentina. *The World Bank Economic Review*, 34(2):393–417.

Altinok, N., Angrist, N., and Patrinos, H. A. (2018). Global data set on education quality (1965-2015). *World Bank Policy Research Working Paper*, (8314).

Angrist, N., Djankov, S., Goldberg, P. K., and Patrinos, H. A. (2021). Measuring human capital using global learning data. *Nature*, 592(7854):403–408.

Angrist, N., Evans, D. K., Filmer, D., Glennerster, R., Rogers, H., and Sabarwal, S. (2025). How to improve education outcomes most efficiently? a review of the evidence using a unified metric. *Journal of Development Economics*, 172:103382.

Ashraf, N., Banerjee, A., and Nourani, V. (2020). Learning to teach by learning to learn. Technical report, Unpublished Working Paper. Job Market Paper.

Baker, F. (2017). The basics of item response theory using r.

Bando, R. and Li, X. (2014). The effect of in-service teacher training on student learning of english as a second language.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M., and Walton, M. (2016). Mainstreaming an effective intervention: Evidence from randomized evaluations of "teaching at the right level" in india. Technical report, National Bureau of Economic Research.

Bassi, M., Meghir, C., and Reynoso, A. (2020). Education quality and teaching practices. *The Economic Journal*, 130(631):1937–1965.

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

Bietenbeck, J., Piopiunik, M., and Wiederhold, S. (2018). Africa's skill tragedy: Does teachers' lack of knowledge lead to low student performance? *Journal of human resources*, 53(3):553–578.

Bold, T., Filmer, D., Martin, G., Molina, E., Rockmore, C., Stacy, B., Svensson, J., and Wane, W. (2017a). What do teachers know and do? does it matter? evidence from primary schools in africa. *Does it Matter*.

Bold, T., Filmer, D., Martin, G., Molina, E., Stacy, B., Rockmore, C., Svensson, J., and Wane, W. (2017b). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in africa. *Journal of Economic Perspectives*, 31(4):185–204.

Bold, T., Filmer, D., Molina, E., and Svensson, J. (2019). The lost human capital: Teacher knowledge and student achievement in africa. *World Bank Policy Research Working Paper*, (8849).

Brunetti, A., Büchel, K., Jakob, M., Jann, B., and Steffen, D. (2024). Inadequate teacher content knowledge and what could be done about it: evidence from el salvador. *Journal of Development Effectiveness*, 16(2):206–229.

Bruns, B., Costa, L., and Cunha, N. (2018). Through the looking glass: Can classroom observation and coaching improve teacher performance in brazil? *Economics of Education Review*, 64:214–250.

Büchel, K., Jakob, M., Kühnhanss, C., Steffen, D., and Brunetti, A. (2022). The relative effectiveness of teachers and learning software: Evidence from a field experiment in el salvador. *Journal of labor economics*, 40(3):737–777.

Buhl-Wiggers, J., Kerwin, J. T., de la Piedra, R. M., Smith, J., and Thornton, R. (2023). Reading for life: Lasting impacts of a literacy intervention in uganda.

Cilliers, J., Fleisch, B., Prinsloo, C., and Taylor, S. (2020). How to improve teaching practice?: an experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*, 55(3):926–962.

Filmer, D. P. and Rogers, F. H. (2019). World development report 2018: Learning to realize education's promise.

Grønmo, L. S., Lindquist, M., Arora, A., and Mullis, I. V. (2015). Timss 2015 mathematics framework. *Timss*, 11:28.

Jukes, M. C., Turner, E. L., Dubeck, M. M., Halliday, K. E., Inyega, H. N., Wolf, S., Zuilkowski, S. S., and Brooker, S. J. (2017). Improving literacy instruction in kenya through teacher professional development and text messages support: A cluster randomized trial. *Journal of Research on Educational Effectiveness*, 10(3):449–481.

Kerwin, J. T. and Thornton, R. L. (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.

Kim, S. (2006). A comparative study of irt fixed parameter calibration methods. *Journal of educational measurement*, 43(4):355–381.

Lee, W.-C. and Lee, G. (2018). Irt linking and equating. *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, pages 639–673.

Loyalka, P., Popova, A., Li, G., and Shi, Z. (2019). Does teacher training actually work? evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics*, 11(3):128–154.

Maruyama, T. and Kurosaki, T. (2024). Developing textbooks to improve math learning in primary education: Empirical evidence from el salvador. *Economic Development and Cultural Change*, 72(2):833–866.

Metzler, J. and Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of development economics*, 99(2):486–496.

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

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., and Fishbein, B. (2020). Timss 2019 international results in mathematics and science. TIMSS & PIRLS International Study Center, Boston College. Accessed: 2025-02-07.

Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, 109(4):1426–1460.

Özler, B., Fernald, L. C., Kariger, P., McConnell, C., Neuman, M., and Fraga, E. (2018). Combining pre-school teacher training with parenting education: A cluster-randomized controlled trial. *Journal of Development Economics*, 133:448–467.

Patel, D. and Sandefur, J. (2020). A rosetta stone for human capital. Technical report.

Piper, B., Zuilkowski, S. S., Dubeck, M., Jepkemei, E., and King, S. J. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development*, 106:324–336.

Popova, A., Evans, D. K., Breeding, M. E., and Arancibia, V. (2018). Teacher professional development around the world. *World Bank Paper Series*, 65.

Popova, A., Evans, D. K., Breeding, M. E., and Arancibia, V. (2022). Teacher professional development around the world: The gap between evidence and practice. *The World Bank Research Observer*, 37(1):107–136.

Schaffner, J., Glewwe, P., and Sharma, U. (2025). Why programs fail: Lessons for improving public service quality from a mixed-methods evaluation of an unsuccessful teacher training program in nepal. *The World Bank Economic Review*, 39(2):473–496.

Sinha, S., Banerji, R., and Wadhwa, W. (2016). *Teacher performance in Bihar, India: Implications for education.* World Bank Publications.

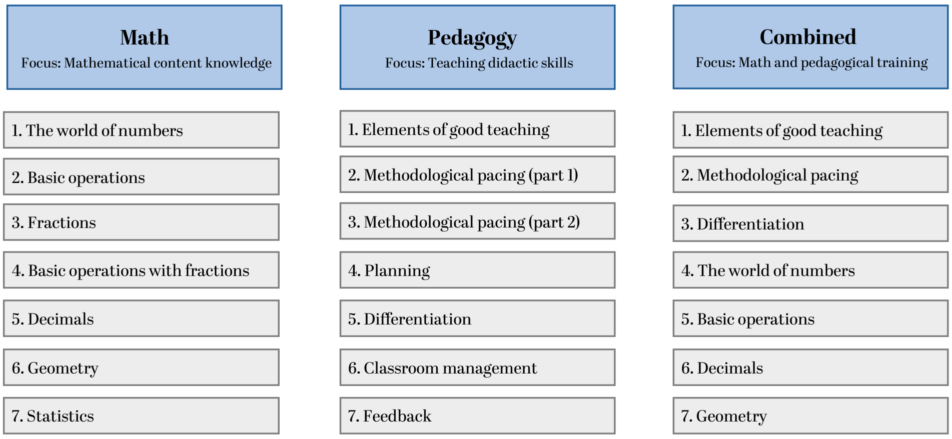
Stallings, J. A. (1973). Follow through program classroom observation evaluation 1971-72.

Wolf, S. (2018). Impacts of pre-service training and coaching on kindergarten quality and student learning outcomes in ghana. *Studies in Educational Evaluation*, 59:112–123.

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M., Weiland, C., Gomez, C. J., Moreno, L., Rolla, A., D'Sa, N., et al. (2015). Experimental impacts of a teacher professional development program in chile on preschool classroom quality and child outcomes. *Developmental psychology*, 51(3):309.

A Appendix

A1 Intervention Design



Math Focus: Mathematical content knowledge	Pedagogy Focus: Teaching didactic skills	Combined Focus: Math and pedagogical training
1. The world of numbers	1. Elements of good teaching	1. Elements of good teaching
2. Basic operations	2. Methodological pacing (part 1)	2. Methodological pacing
3. Fractions	3. Methodological pacing (part 2)	3. Differentiation
4. Basic operations with fractions	4. Planning	4. The world of numbers
5. Decimals	5. Differentiation	5. Basic operations
6. Geometry	6. Classroom management	6. Decimals
7. Statistics	7. Feedback	7. Geometry

Figure A1: Intervention training block topics

Each training block corresponded to 3 weeks, including one in-person workshop, self-study modules and personalized coaching. See figure 2 for an additional reference.

A2 Data Processing

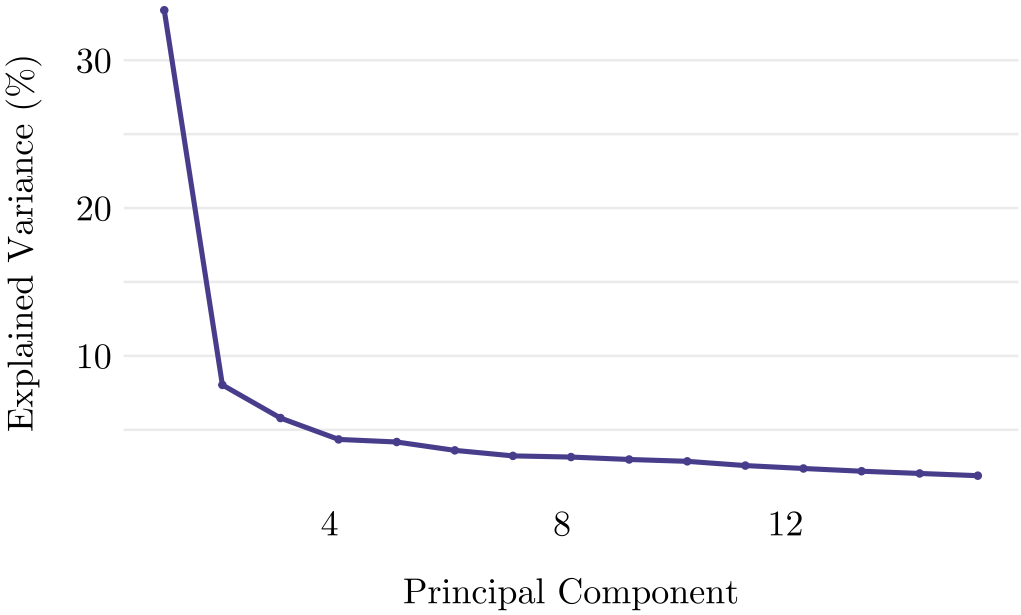


Figure A2: Scree Plot

The Scree plot shows the explained variance for each principal component in a factor analysis of all pedagogy evaluation criteria used for the pedagogical teacher assessment at follow-up.

A3 Compliance and Attrition

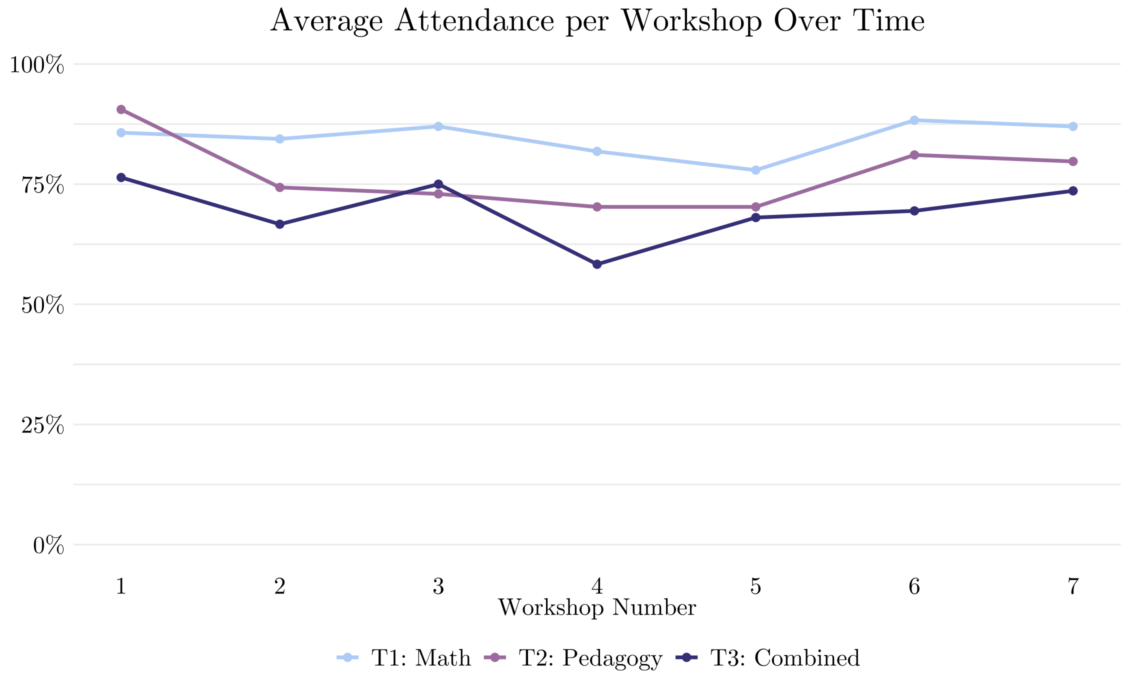


Figure A3: Compliance over Time

Table A1: Attritor Characteristics: Teacher Tests

	Control	T1: Math	T2: Pedagogy	T3: Combined	P-value	N
Endline						
Baseline score	0.53	0.64	0.50	0.65	0.21	29
Age	39.60	50.00	47.89	43.33	0.50	28
Sex	0.70	0.50	0.78	0.67	0.36	29
Follow-up						
Baseline score	0.52	0.61	0.55	0.51	1.00	41
Age	39.62	53.00	45.92	46.33	0.23	39
Sex	0.62	0.62	0.75	0.77	0.78	41

The last row indicates the p-value of a joint F-Test that each treatment dummy coefficient is equal to 0. Balance is calculated using strata fixed effects. Standard errors are clustered at the teacher level.

Table A2: Attritor Characteristics: Student Tests

	Control	T1: Math	T2: Pedagogy	T3: Combined	P-value	N
Endline						
Baseline score	0.24	0.27	0.25	0.25	0.95	1060
Age	11.20	10.76	10.99	10.99	0.34	2079
Sex	0.47	0.48	0.47	0.48	0.93	2234
SES	0.69	0.74	0.70	0.71	0.08	2069
Follow-up						
Baseline score	0.28	0.28	0.27	0.29	0.12	3569
Age	10.68	10.76	10.71	10.71	0.67	3351
Sex	0.49	0.49	0.52	0.48	0.47	3546
SES	0.69	0.71	0.71	0.70	0.40	3272

The last row indicates the p-value of a joint F-Test that each treatment dummy coefficient is equal to 0. Balance is calculated using strata fixed effects. Standard errors are clustered at the teacher level.

Table A3: Attritor Characteristics: Classroom Observations

	Control	T1: Math	T2: Pedagogy	T3: Combined	P-value	N
Endline						
Baseline score	0.57	0.47	0.57	0.67	0.02	29
Age	39.88	45.40	48.50	39.40	0.73	29
Sex	0.50	1.00	0.50	0.70	0.59	29
Follow-up						
Baseline score	0.56	0.56	0.54	0.51	0.08	107
Age	43.04	47.84	47.42	45.93	0.45	106
Sex	0.68	0.80	0.62	0.75	0.26	107

The last row indicates the p-value of a joint F-Test that each treatment dummy coefficient is equal to 0. Balance is calculated using strata fixed effects. Standard errors are clustered at the teacher level.

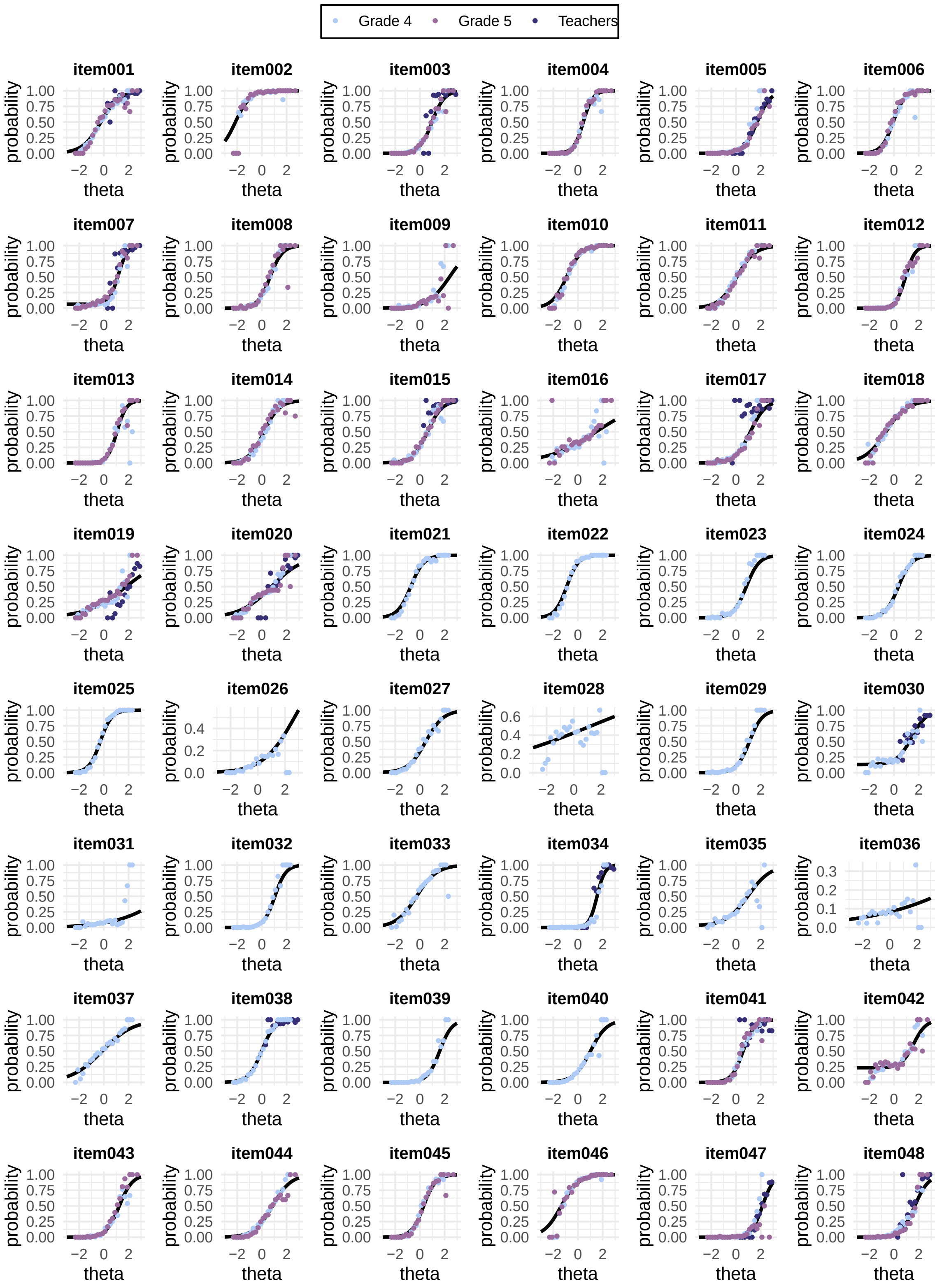
Table A4: Attrition Rates per Experimental Groups

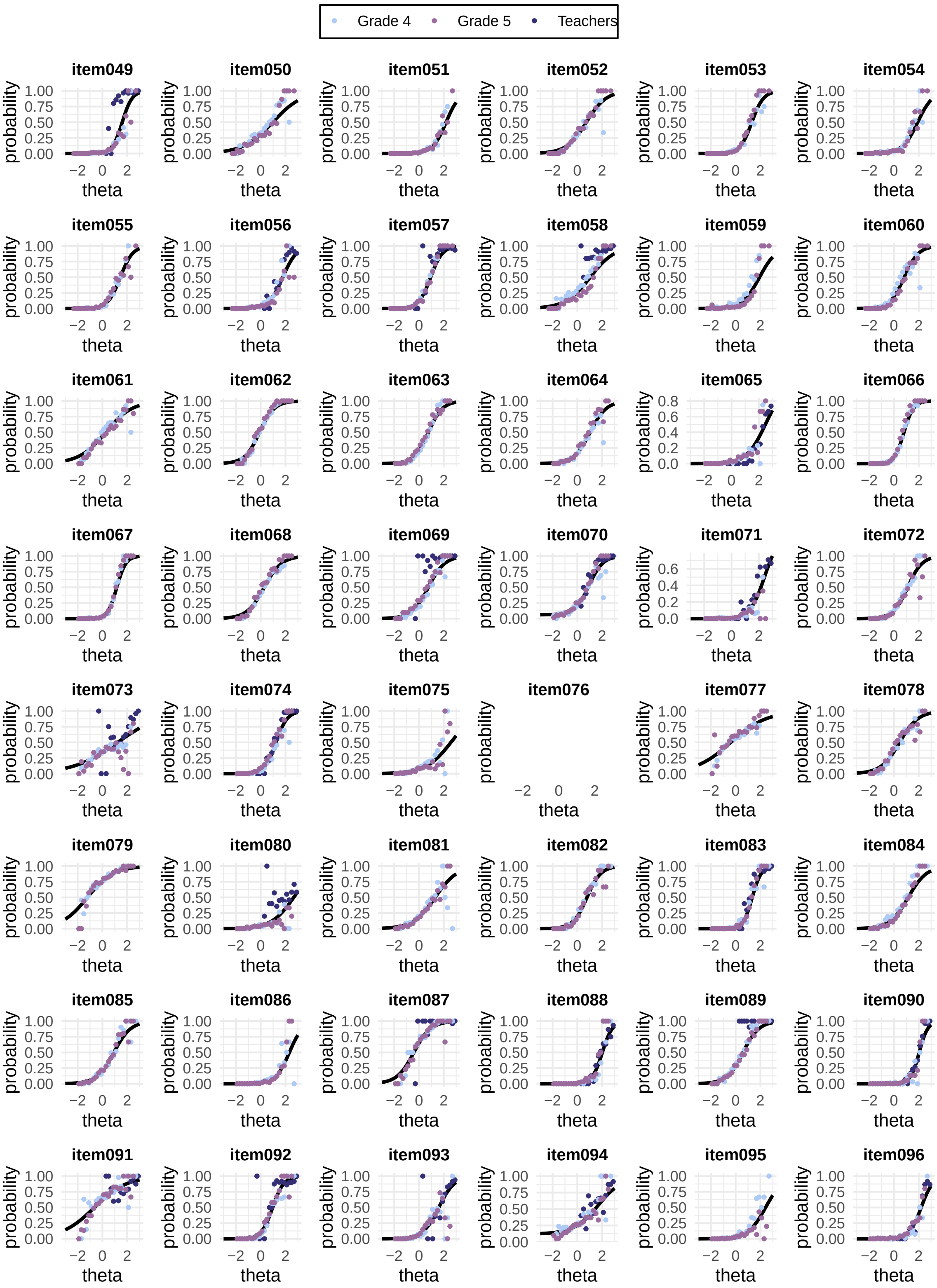
	Control	T1: Math	T2: Pedagogy	T3: Combined	P-value	N
Endline						
Teacher tests	0.12	0.05	0.11	0.07	0.32	338
Classroom observations	0.10	0.06	0.07	0.12	0.51	338
Student tests	0.16	0.17	0.19	0.19	0.45	6011
Follow-up						
Teacher tests	0.10	0.09	0.14	0.15	0.50	338
Classroom observations	0.33	0.29	0.31	0.33	0.93	338
Student tests	0.59	0.55	0.55	0.70	0.09	6011

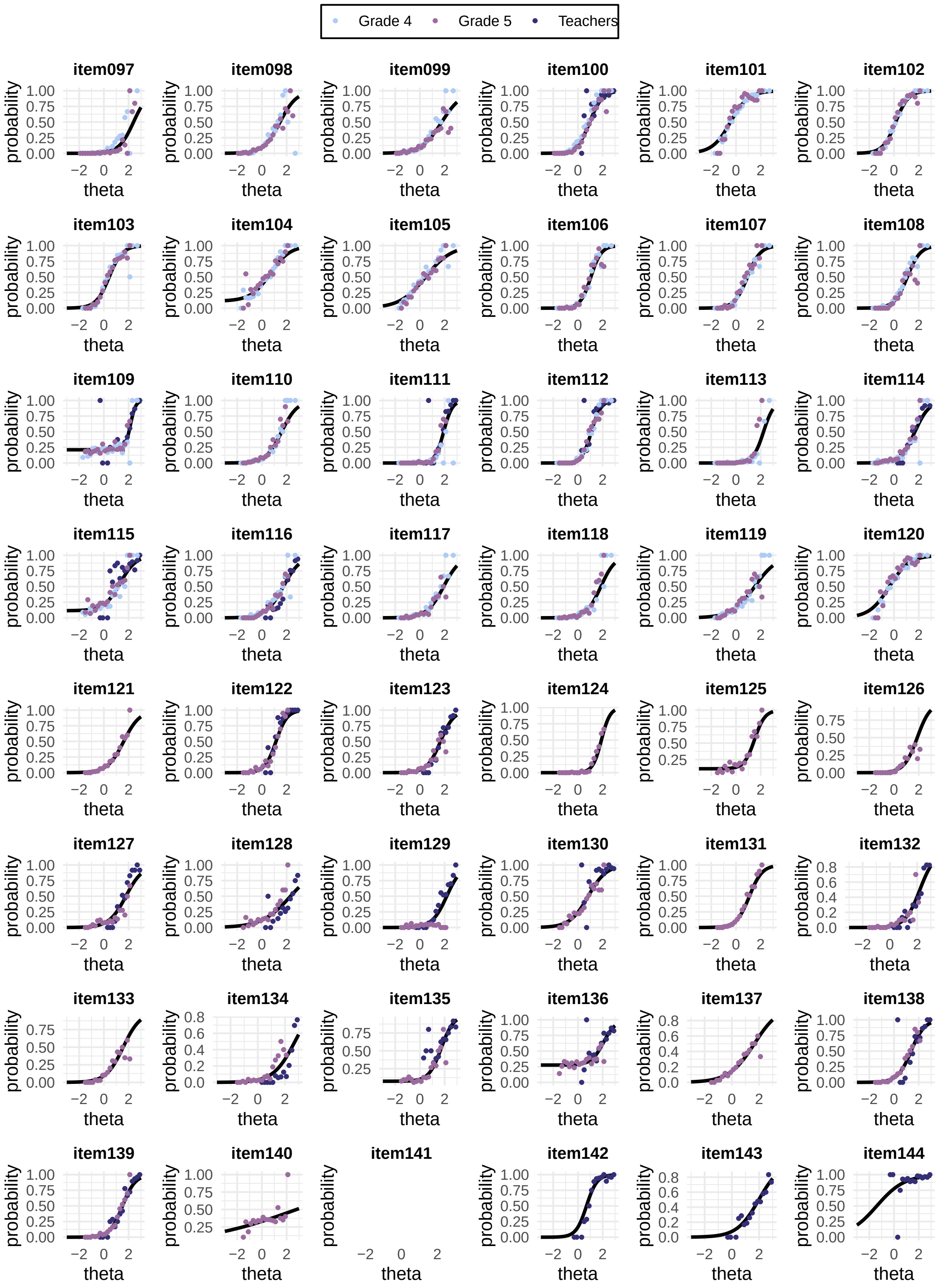
The last row indicates the p-value of a joint F-Test that each treatment dummy coefficient is equal to 0. Balance is calculated using strata fixed effects. Standard errors are clustered at the teacher level.

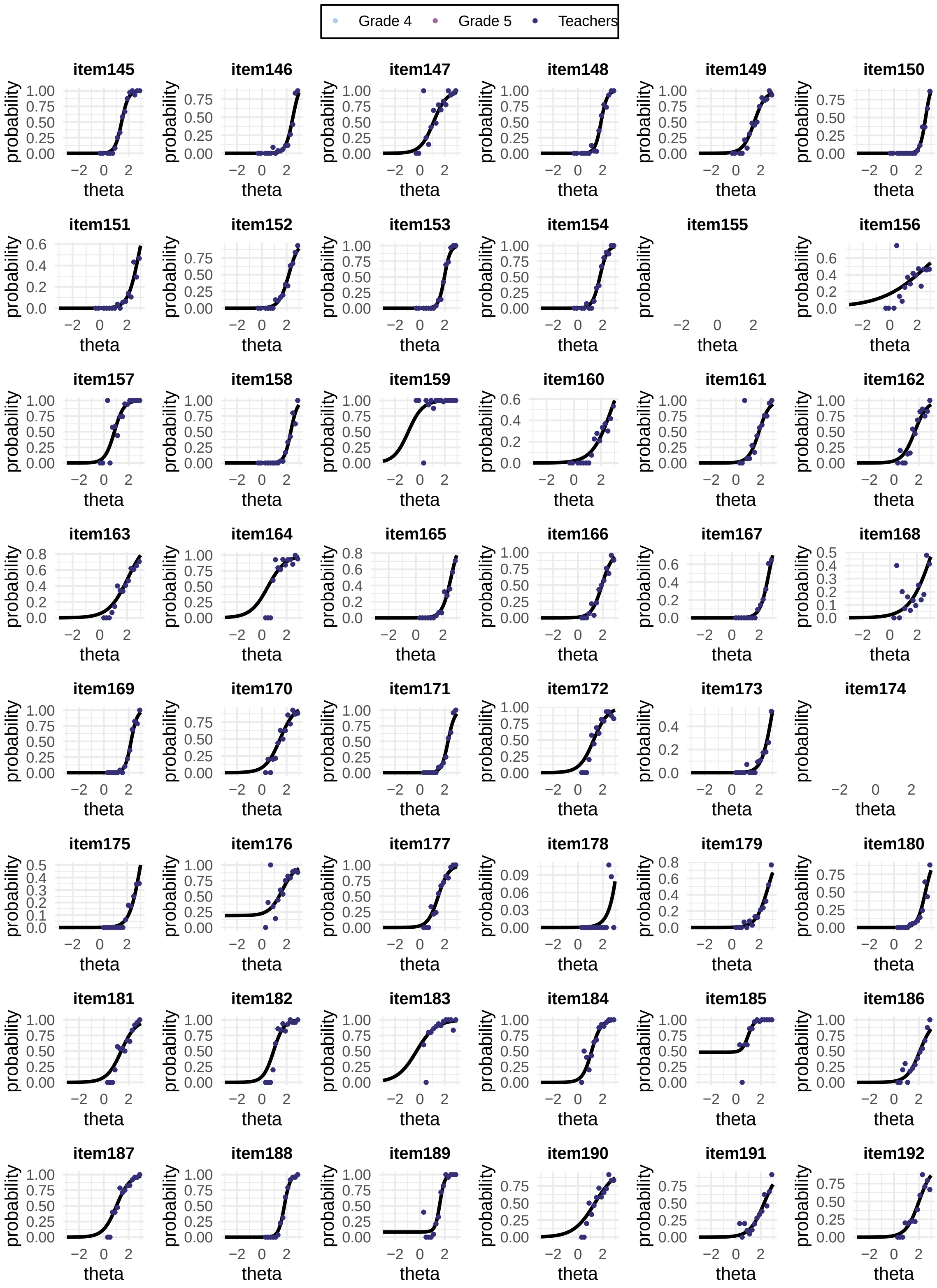
A4 Item Response Theory

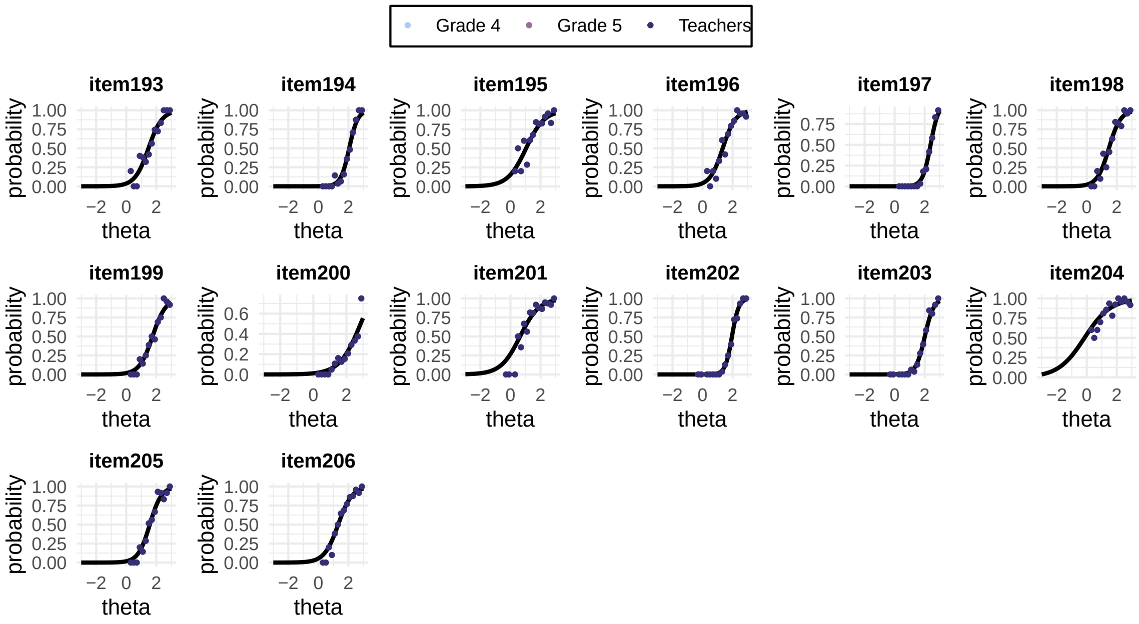
Figure A4: DIF plots




● Grade 4
● Grade 5
● Teachers
item049
item050
item051
item052
item053
item054
item055
item056
item057
item058
item059
item060
item061
item062
item063
item064
item065
item066
item067
item068
item069
item070
item071
item072
item073
item074
item075
item076
item077
item078
item079
item080
item081
item082
item083
item084
item085
item086
item087
item088
item089
item090
item091
item092
item093
item094
item095
item096
probability
1.00
0.75
0.50
0.25
0.00
0.8
0.6
0.4
0.2
0.0
−2
0
2
theta


Grade 4
Grade 5
Teachers
item097
item098
item099
item100
item101
item102
item103
item104
item105
item106
item107
item108
item109
item110
item111
item112
item113
item114
item115
item116
item117
item118
item119
item120
item121
item122
item123
item124
item125
item126
item127
item128
item129
item130
item131
item132
item133
item134
item135
item136
item137
item138
item139
item140
item141
item142
item143
item144
probability
theta
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8
−2
0
2


Grade 4
Grade 5
Teachers
item145
item146
item147
item148
item149
item150
item151
item152
item153
item154
item155
item156
item157
item158
item159
item160
item161
item162
item163
item164
item165
item166
item167
item168
item169
item170
item171
item172
item173
item174
item175
item176
item177
item178
item179
item180
item181
item182
item183
item184
item185
item186
item187
item188
item189
item190
item191
item192
probability
theta
−2
0
2



● Grade 4
● Grade 5
● Teachers
item193
item194
item195
item196
item197
item198
item199
item200
item201
item202
item203
item204
item205
item206
probability
theta
1.00
0.75
0.50
0.25
0.00
0.6
0.4
0.2
0.0
−2
0
2

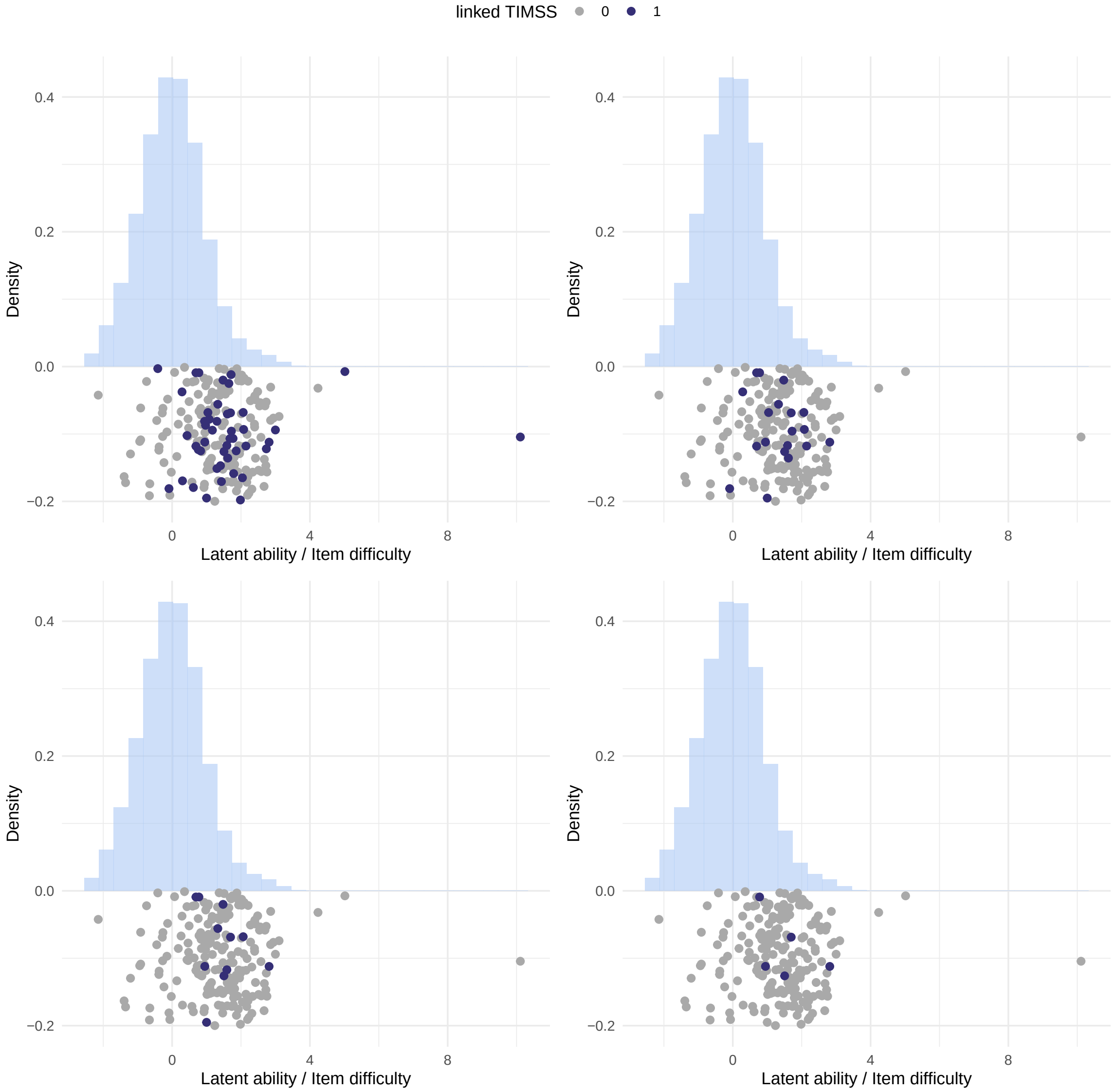
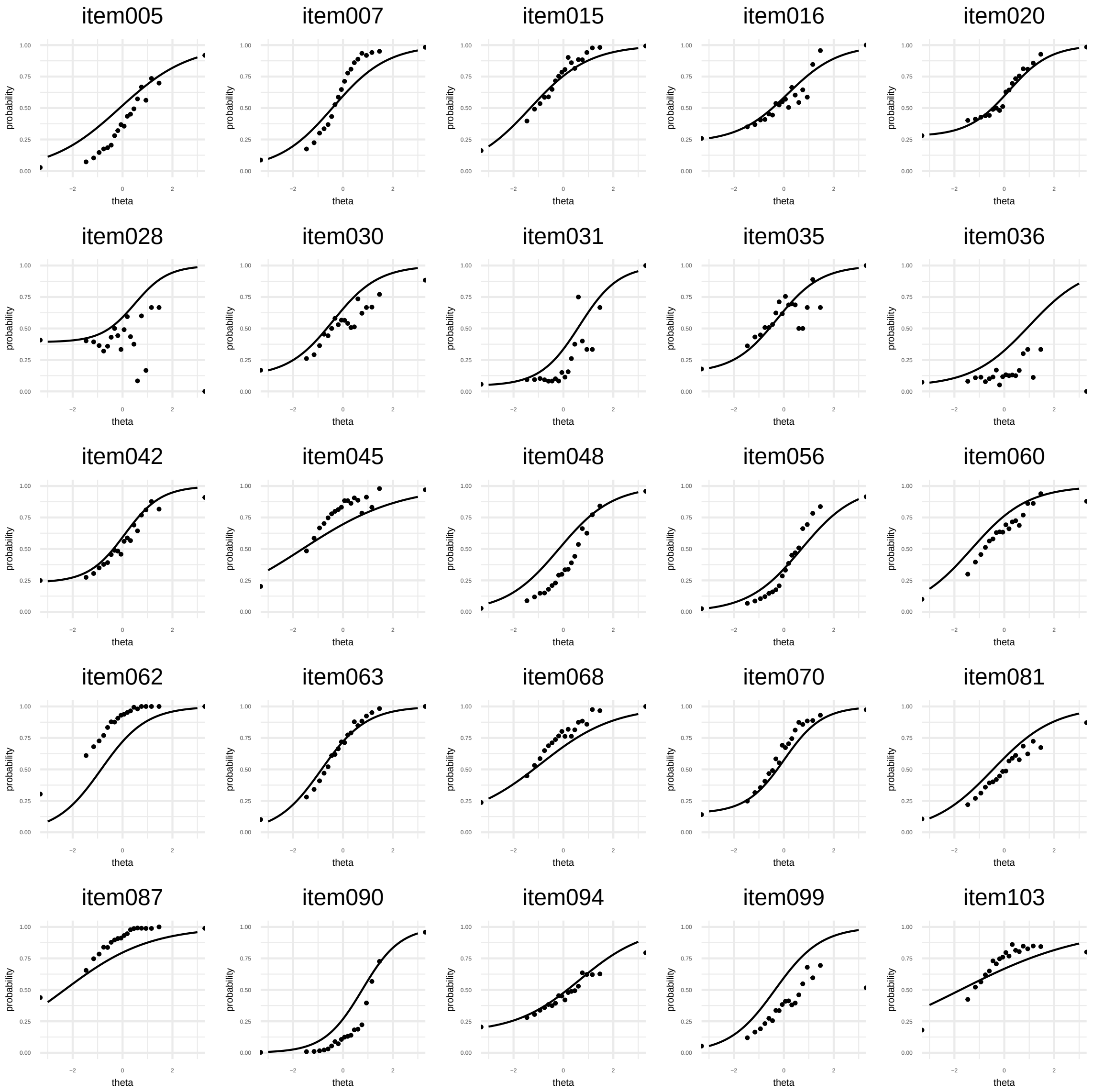
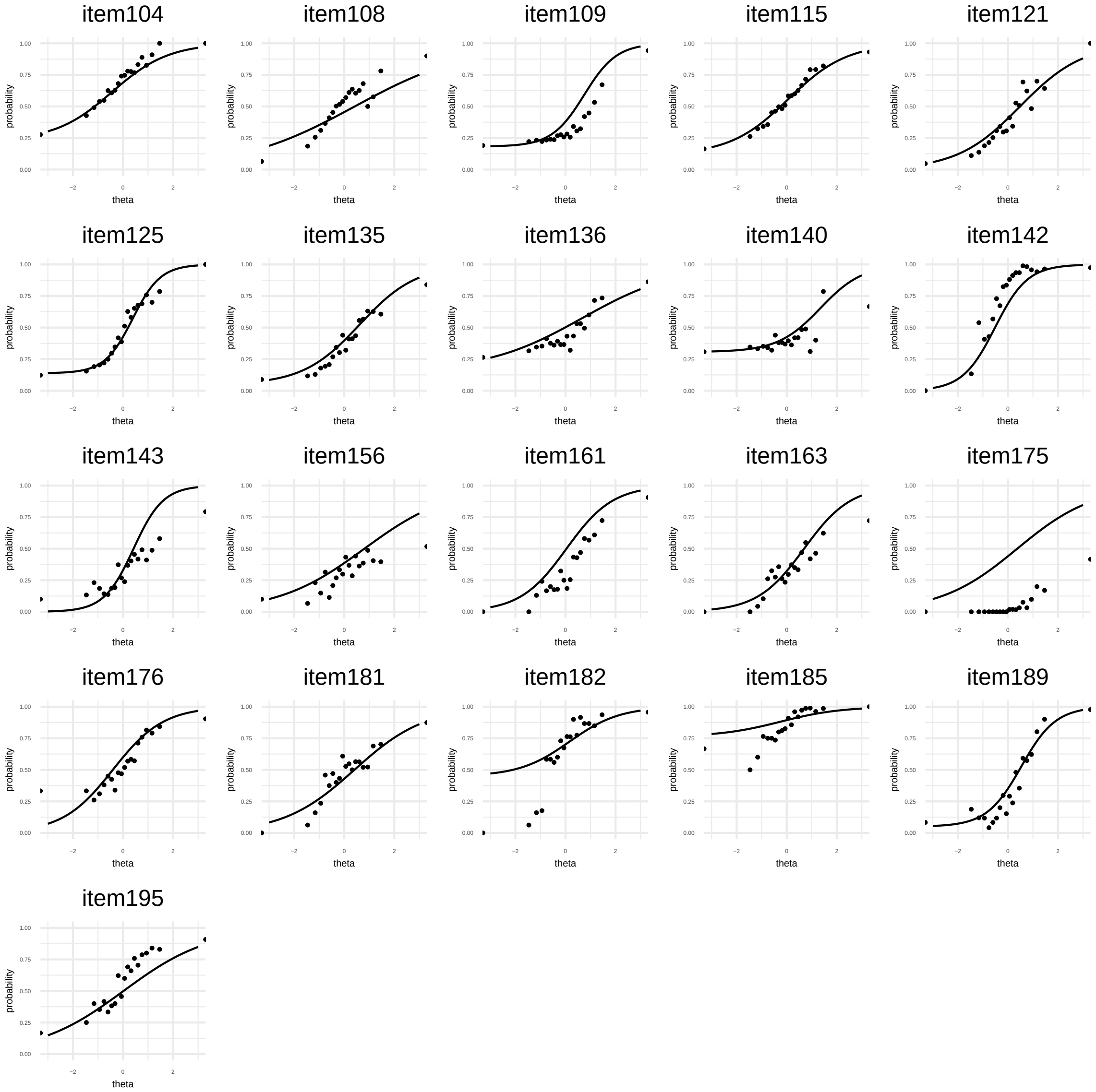


Figure A5: Wright Map

The Wright Map, or person–item map, displays the distribution of the latent trait in the population (top part of the figure) alongside the distribution of item difficulties in the test (bottom part of the figure). The top left panel highlights all TIMSS items, while the top right restricts attention to TIMSS items without substantial DIF. The bottom left panel highlights only items with very little or no DIF, and the bottom right highlights those with no DIF. All difficulty parameters are taken from the freely estimated model in Section 2.4.1, which does not impose TIMSS linking.

Figure A6: DIF plots for TIMSS linking (anchor items)




item104
item108
item109
item115
item121
item125
item135
item136
item140
item142
item143
item156
item161
item163
item175
item176
item181
item182
item185
item189
item195
probability
theta
1.00
0.75
0.50
0.25
0.00
−2
0
2

A5 Survey and Classroom Observation Data

Table A5: Survey outcomes: Teachers

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Workload	0.420**	0.479***	0.544***	307	0.228	0.174	0.149	293
	(0.166)	(0.171)	(0.168)		(0.177)	(0.182)	(0.179)	
Teacher Impact Index	0.108	-0.147	0.071	308	-0.119	-0.309*	-0.056	296
	(0.163)	(0.168)	(0.165)		(0.171)	(0.175)	(0.174)	
Job Satisfaction Index	-0.199	0.013	0.150	305	-0.070	0.041	-0.077	293
	(0.149)	(0.153)	(0.151)		(0.153)	(0.158)	(0.155)	
Didactics Index	0.155	0.066	0.003	308	0.092	0.100	0.083	296
	(0.156)	(0.161)	(0.158)		(0.165)	(0.171)	(0.169)	
Preparation Index	0.084	0.019	-0.057	308	0.090	0.013	-0.069	296
	(0.153)	(0.158)	(0.155)		(0.143)	(0.147)	(0.146)	

All the dependent variables are standardized unless stated otherwise. Controls include strata fixed effects and demographic teacher variables. Standard errors are clustered at the teacher level.

Table A6: Survey outcomes: Teachers

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Workload	0.420**	0.479***	0.544***	307	0.228	0.174	0.149	293
	(0.166)	(0.171)	(0.168)		(0.177)	(0.182)	(0.179)	
Perceived impact on students								
Teacher	-0.007	-0.212	0.222	307	-0.170	-0.307*	-0.155	296
	(0.168)	(0.172)	(0.169)		(0.173)	(0.178)	(0.176)	
Didactics knowledge	0.084	-0.089	0.062	308	0.008	-0.169	0.124	296
	(0.168)	(0.173)	(0.169)		(0.173)	(0.177)	(0.176)	
Math knowledge	0.147	-0.026	-0.107	308	-0.123	-0.269	-0.099	296
	(0.157)	(0.161)	(0.159)		(0.173)	(0.177)	(0.176)	
Lesson preparation								
Time (in minutes)	-0.105	0.001	0.022	301	0.152	-0.069	-0.059	277
	(0.157)	(0.161)	(0.158)		(0.178)	(0.182)	(0.182)	
Esmate usage	0.088	0.095	0.003	306	0.130	0.113	0.014	295
	(0.116)	(0.120)	(0.118)		(0.102)	(0.105)	(0.104)	
Own planning	0.063	-0.049	-0.212	300	-0.100	-0.045	-0.305*	290
	(0.160)	(0.165)	(0.164)		(0.170)	(0.174)	(0.171)	
Own activities	-0.035	0.002	-0.156	306	0.010	0.003	-0.054	292
	(0.147)	(0.151)	(0.149)		(0.155)	(0.161)	(0.158)	
Own didactic material	0.416***	0.220	0.107	306	0.283*	0.212	0.029	292
	(0.158)	(0.163)	(0.160)		(0.149)	(0.154)	(0.152)	
Homework check	0.083	-0.219	-0.036	304	-0.101	-0.157	-0.018	291
	(0.168)	(0.172)	(0.170)		(0.204)	(0.212)	(0.209)	
No Preparation	-0.095	0.058	-0.016	286	-0.026	-0.028	-0.156	271
	(0.152)	(0.159)	(0.157)		(0.153)	(0.157)	(0.156)	
During the lesson								
Satisfaction	-0.120	-0.384*	-0.322	304	-0.121	-0.291	0.027	293
	(0.215)	(0.221)	(0.219)		(0.172)	(0.177)	(0.174)	
Frustration	-0.203	0.028	0.126	303	0.020	0.062	-0.040	291
	(0.160)	(0.164)	(0.162)		(0.153)	(0.159)	(0.155)	
Methodology variation	0.069	0.179	0.002	303	0.064	0.201	0.175	294
	(0.154)	(0.158)	(0.156)		(0.171)	(0.175)	(0.174)	
Example usage	0.079	0.212	0.109	304	0.046	0.118	0.201	293
	(0.161)	(0.165)	(0.163)		(0.151)	(0.155)	(0.153)	
Difficulity didactics knowledge	-0.305**	-0.189	-0.073	304	-0.183	-0.180	-0.049	291
	(0.138)	(0.142)	(0.141)		(0.140)	(0.145)	(0.141)	
Difficultiy math knowledge	-0.113	-0.082	0.099	305	-0.109	-0.050	-0.220	291
	(0.147)	(0.151)	(0.150)		(0.150)	(0.154)	(0.151)	
Comparison to other subjects								
Time (in minutes)	-5.032	-5.749	-2.419	264	5.775	-2.432	4.393	251
	(7.021)	(7.192)	(6.891)		(9.655)	(10.039)	(9.723)	
Freedom	0.035	-0.127	0.110	264	-0.192	-0.033	0.109	262
	(0.166)	(0.169)	(0.166)		(0.172)	(0.176)	(0.174)	
Student interest	-0.046	0.178	0.270*	266	-0.137	-0.177	-0.266	265
	(0.154)	(0.156)	(0.152)		(0.169)	(0.172)	(0.169)	

Table A7: Survey outcomes: Students

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Favorite subject is math	0.036	0.012	-0.011	4321	-0.020	-0.103**	-0.048	2109
	(0.026)	(0.027)	(0.026)		(0.048)	(0.047)	(0.046)	
Least favorite subject is math	-0.035	0.012	0.006	3923	0.039	0.088*	0.040	1939
	(0.027)	(0.025)	(0.026)		(0.041)	(0.047)	(0.044)	
Favorite activity is school	0.026	-0.032	-0.001	4360	0.036	0.001	-0.016	2165
	(0.024)	(0.028)	(0.026)		(0.034)	(0.035)	(0.041)	
Least favorite activity is school	-0.010	0.018	0.016	4032	-0.013	0.010	-0.011	1976
	(0.015)	(0.018)	(0.017)		(0.023)	(0.024)	(0.027)	

All the dependent variables are standardized. Controls include strata fixed effects and demographic teacher variables. Standard errors are clustered at the teacher level.

Table A8: Classroom observations extended: Teachers

	Endline			*Follow-up*		
	T1: Math	T2: Ped	T3: Both	T1: Math	T2: Ped	T3: Both
Teacher use of time						
Active (%)	-0.013	-0.002	0.014	-0.014	-0.052	0.027
	(0.036)	(0.036)	(0.036)	(0.039)	(0.039)	(0.039)
Passive (%)	0.006	0.035	-0.017	0.028	0.050	0.001
	(0.035)	(0.036)	(0.036)	(0.033)	(0.033)	(0.033)
Management (%)	0.014	-0.018	0.010	0.035	0.044	0.005
	(0.024)	(0.024)	(0.025)	(0.028)	(0.028)	(0.028)
Social (%)	-0.006	-0.015	-0.007	-0.028	-0.032*	-0.020
	(0.015)	(0.015)	(0.015)	(0.018)	(0.018)	(0.018)
Material use						
Good Material (%)	0.039*	0.051**	0.047*	0.034	0.022	0.030
	(0.023)	(0.024)	(0.024)	(0.025)	(0.025)	(0.025)
Bad/Non-interactive Material (%)	-0.057	-0.078**	-0.082**	-0.061	-0.038	-0.028
	(0.038)	(0.038)	(0.039)	(0.040)	(0.041)	(0.041)
Material variation (sd)	0.104	0.189	0.059	-0.006	-0.134	0.120
	(0.147)	(0.148)	(0.150)	(0.170)	(0.172)	(0.172)
Pedagogy						
Does lesson outline (sd)	0.135	0.208	0.251	-0.014	0.160	-0.120
	(0.160)	(0.161)	(0.163)	(0.157)	(0.160)	(0.159)
Does homework review (sd)	-0.130	-0.021	0.017	0.011	0.019	-0.012
	(0.164)	(0.165)	(0.167)	(0.170)	(0.173)	(0.171)
Uses examples (sd)	0.191	0.185	0.217	-0.280	-0.150	-0.029
	(0.155)	(0.156)	(0.158)	(0.190)	(0.192)	(0.192)
Moves in classroom (sd)	0.018	0.138	0.145	0.113	0.278*	0.216
	(0.153)	(0.154)	(0.155)	(0.167)	(0.168)	(0.168)
Gives homework (sd)	0.117	0.329**	0.260*	-0.186	0.093	-0.258
	(0.148)	(0.149)	(0.151)	(0.169)	(0.170)	(0.170)
Does lesson summary (sd)	0.145	0.292*	0.278*	-0.191	-0.028	-0.174
	(0.160)	(0.160)	(0.163)	(0.178)	(0.179)	(0.179)
Activities						
Activity variation (sd)	0.200	-0.025	0.059	0.173	-0.070	0.072
	(0.141)	(0.142)	(0.144)	(0.161)	(0.162)	(0.162)
Conducts interactive activity	0.010	0.184***	0.128**	0.107**	0.123**	0.050
	(0.058)	(0.058)	(0.059)	(0.051)	(0.051)	(0.051)
Distracted students						
One or more students distracted (%)	-0.010	-0.090*	-0.048	-0.007	-0.061	-0.023
	(0.046)	(0.047)	(0.047)	(0.047)	(0.047)	(0.047)
Group distracted (%)	-0.010	-0.077*	-0.054	0.028	-0.026	-0.011
	(0.045)	(0.045)	(0.046)	(0.053)	(0.054)	(0.054)

Sample sizes are n = 307 for the estimation of endline effects and n = 230 for follow-up effects. All the dependent variables are either standardized (sd), binary or indicated as a percentage of the lesson time (%). Good material constitutes shared material, learning aids or ICT. Bad/Non-interactive material consitutes textbook, notebook

A6 Main Effects

Table A9: ITT estimates math scores: IRT

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Teachers								
IRT math score	0.139*	-0.144*	0.088	308	0.291***	0.063	0.136	297
	(0.074)	(0.076)	(0.075)		(0.082)	(0.085)	(0.084)	
Students								
IRT math score	0.010	-0.042	-0.096	4950	0.014	-0.043	-0.029	2441
	(0.062)	(0.072)	(0.061)		(0.078)	(0.086)	(0.095)	

All the dependent variables are standardized. Controls include strata fixed effects, the outcome variable at baseline and demographic controls. Results for teachers include demographic controls. Standard errors are clustered at the teacher level. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Table A10: ITT estimates: Results per grade

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Standardized scores								
4th grade	0.168*	0.101	-0.051	2443	0.011	0.008	-0.085	1188
	(0.088)	(0.098)	(0.086)		(0.113)	(0.121)	(0.148)	
5th grade	-0.136	-0.121	-0.164*	2507	-0.011	-0.009	-0.074	1244
	(0.084)	(0.095)	(0.086)		(0.108)	(0.098)	(0.122)	
IRT scores								
4th grade	0.136*	0.041	-0.035	2443	0.010	-0.055	-0.129	1188
	(0.082)	(0.095)	(0.084)		(0.104)	(0.111)	(0.139)	
5th grade	-0.148**	-0.148*	-0.172**	2507	0.012	-0.011	-0.012	1244
	(0.076)	(0.087)	(0.081)		(0.103)	(0.104)	(0.116)	

All the dependent variables are standardized. Controls include strata fixed effects and the outcome variable at baseline. Standard errors are clustered at the teacher level. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Table A14: Teacher content knowledge and student learning gains: Evidence from quasi-experimental data

Study	Student Effect (+ 1σ teacher score)	Country/Region	Grade	Empirical Strategy
Metzler and Woessmann (2012)	Math: 0.09, Language: 0.03 (insig.)	Peru	Grade 6	Teacher FE + Student FE
Bietenbeck, Piopiunik, and Wiederhold (2018)	Mixed: 0.03	6 East African countries	Grade 6	Teacher FE + Student FE
Bold et al. (2019)	Mixed: 0.07	7 African countries	Grade 4	Teacher FE + Student FE
Bau and Das (2020)	Math: 0.09, Language: 0.06	Pakistan, Punjab	Grade 3-5	Teacher value-added approach
Brunetti et. al (2023)	Math: 0.09–0.12	El Salvador, Morazán	Grade 3- 6	Various controls
Our results	Math: 0..07-0.08	El Salvador, Morazán	Grade 4-6	Various controls

Table A11: ITT estimates pedagogical test: Disagregated

	Follow-up			
	T1: Math	T2: Ped	T3: Both	N
Lesson Plan (part 1)				
Exam completness	0.291*	0.475***	0.296*	297
	(0.157)	(0.161)	(0.160)	
Overall quality	0.422**	0.437**	0.380**	297
	(0.167)	(0.171)	(0.170)	
Pedagogy score	0.612***	0.569***	0.574***	297
	(0.194)	(0.200)	(0.198)	
Structure	0.243*	0.439***	0.289*	297
	(0.145)	(0.149)	(0.148)	
Practical planning	0.340*	0.565***	0.193	297
	(0.181)	(0.186)	(0.185)	
Classroom Management (part 2)				
Number of suggested measures	0.098	0.258	0.080	297
	(0.170)	(0.175)	(0.173)	
Measure apropriateness	0.317*	0.309*	0.317*	295
	(0.182)	(0.187)	(0.185)	

All the dependent variables are standardized. Controls include strata fixed effects, the outcome variable at baseline and demographic controls. Standard errors are clustered at the teacher level. The overall quality evaluates the clarity and effectiveness of the lesson, including introduction, content, practice, and summary. The pedagogy score refers to the use of diverse methods, student differentiation, engagement, and material variety. Practical planning measures definition of objectives, time allocation, and material listing. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Table A12: ITT estimates: Teachers subscores

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Number Sense & Place Value	0.099	-0.100	0.179	308	-0.077	0.030	-0.186	297
	(0.117)	(0.121)	(0.119)		(0.128)	(0.132)	(0.131)	
Addition & Subtraction	0.035	0.061	0.027	308	0.070	-0.015	0.070	297
	(0.144)	(0.148)	(0.146)		(0.143)	(0.147)	(0.146)	
Multiplication & Division	-0.037	-0.182	0.006	308	0.315**	-0.180	0.148	297
	(0.133)	(0.137)	(0.134)		(0.146)	(0.151)	(0.150)	
Fractions & Decimals	0.122	-0.128	0.091	308	0.275***	0.070	0.187**	297
	(0.086)	(0.088)	(0.087)		(0.088)	(0.090)	(0.090)	
Algebraic Thinking	0.117	-0.042	-0.103	308	0.171	0.139	0.078	297
	(0.114)	(0.117)	(0.115)		(0.115)	(0.118)	(0.117)	
Order of Operations / Combined Operations	-0.154	-0.040	-0.113	308	-	-	-	-
	(0.147)	(0.151)	(0.149)					
Geometry	0.211**	-0.037	0.249**	308	0.273**	0.140	0.179*	297
	(0.107)	(0.110)	(0.108)		(0.106)	(0.109)	(0.108)	
Measurement & Units	0.264*	0.108	0.150	308	0.119	-0.129	0.154	297
	(0.135)	(0.139)	(0.137)		(0.166)	(0.171)	(0.170)	
Data & Statistics	0.080	-0.127	0.007	308	0.133	-0.152	-0.083	297
	(0.120)	(0.123)	(0.121)		(0.136)	(0.140)	(0.139)	
Proportionality & Percentage	0.082	0.013	-0.059	308	0.177	0.130	0.032	297
	(0.113)	(0.116)	(0.114)		(0.130)	(0.134)	(0.133)	

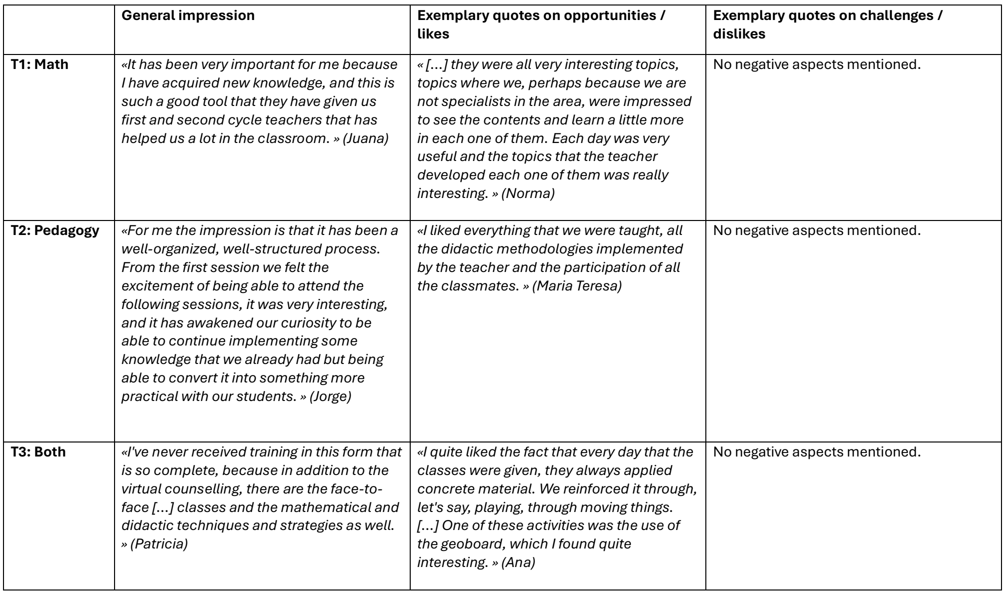
All the dependent variables are standardized unless stated otherwise. Controls include strata fixed effects and demographic teacher variables. Standard errors are clustered at the teacher level.

Table A13: ITT estimates: Students subscores

	Endline				*Follow-up*			
	T1: Math	T2: Ped	T3: Both	N	T1: Math	T2: Ped	T3: Both	N
Number Sense & Place Value	0.037	-0.021	-0.078	4950	0.029	-0.010	-0.013	2441
	(0.059)	(0.061)	(0.061)		(0.084)	(0.084)	(0.091)	
Addition & Subtraction	0.054	0.005	-0.102*	4950	0.043	-0.079	-0.093	2441
	(0.061)	(0.063)	(0.060)		(0.076)	(0.079)	(0.091)	
Multiplication & Division	0.006	0.035	-0.055	4950	-0.040	-0.098	-0.106	2441
	(0.049)	(0.057)	(0.049)		(0.085)	(0.081)	(0.090)	
Fractions & Decimals	0.040	0.068	-0.096	4950	0.028	-0.054	-0.080	2441
	(0.064)	(0.076)	(0.064)		(0.086)	(0.081)	(0.104)	
Algebraic Thinking	-0.002	-0.052	-0.083	4950	-0.012	-0.011	-0.016	2441
	(0.066)	(0.065)	(0.060)		(0.062)	(0.058)	(0.068)	
Order of Operations / Combined Operations	0.089	-0.032	-0.101	4950	0.133	0.230	0.100	1197
	(0.078)	(0.071)	(0.064)		(0.126)	(0.192)	(0.129)	
Geometry	0.008	0.030	-0.003	4950	0.022	0.086	0.107	2441
	(0.069)	(0.079)	(0.069)		(0.092)	(0.089)	(0.120)	
Measurement & Units	0.003	0.032	0.063	4950	-0.011	0.102	0.115	2441
	(0.050)	(0.051)	(0.056)		(0.087)	(0.117)	(0.098)	
Data & Statistics	0.042	0.027	0.025	4950	-0.002	0.088	0.048	2441
	(0.047)	(0.058)	(0.052)		(0.063)	(0.068)	(0.079)	
Proportionality & Percentage	-	-	-	-	-0.122	0.006	-0.036	1244
					(0.084)	(0.090)	(0.112)	

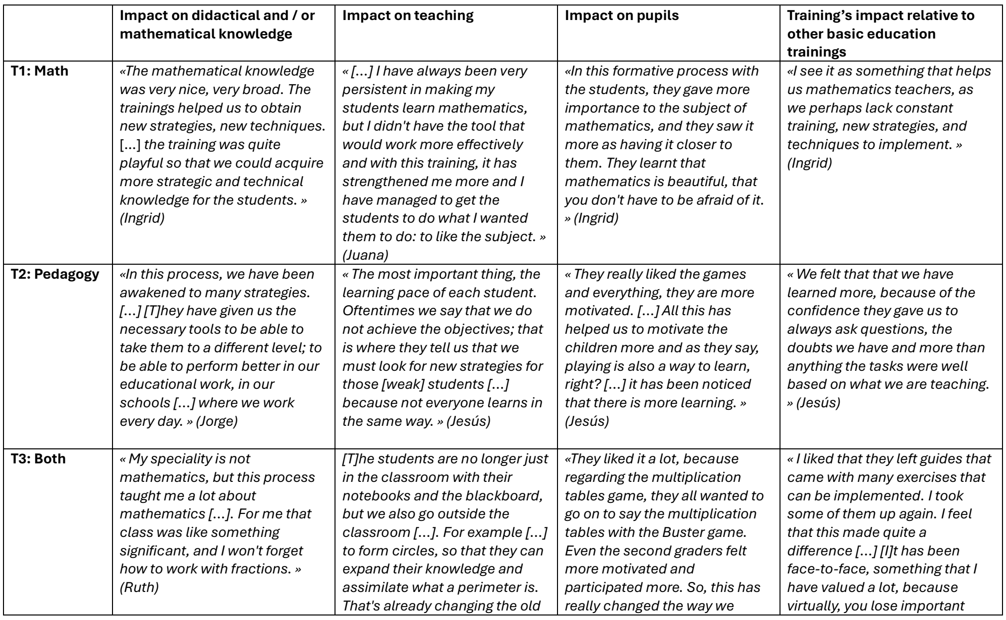
All the dependent variables are standardized. Controls include strata fixed effects and demographic teacher variables. Standard errors are clustered at the teacher level. In the follow-up exams, Data & Statistics items only appeared in grade 5 exams, so data is only available for the younger cohort.

A7 Semi-Structured Interviews with Teachers



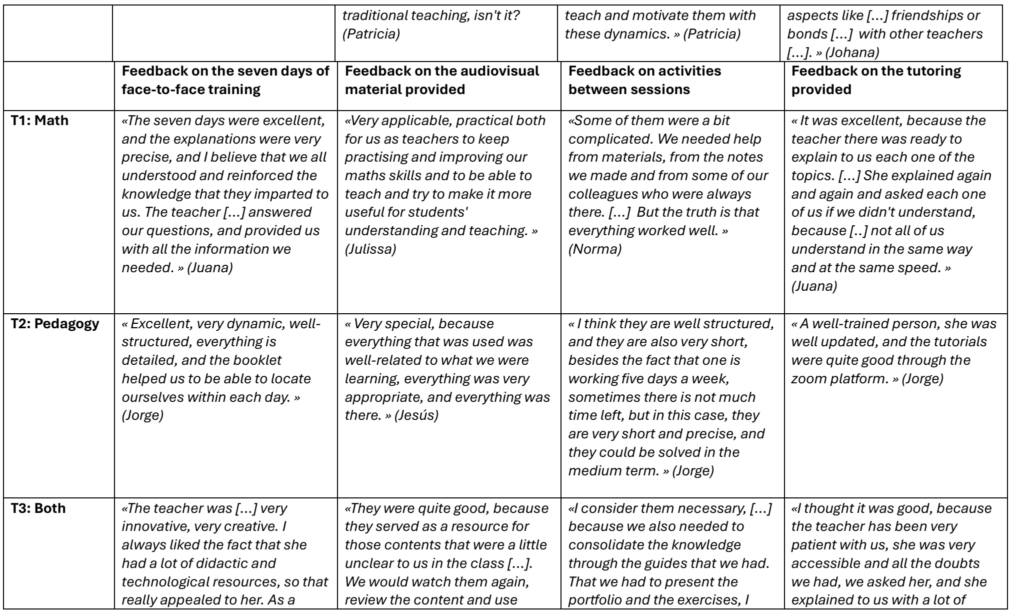
	General impression	Exemplary quotes on opportunities / likes	Exemplary quotes on challenges / dislikes
T1: Math	*«It has been very important for me because I have acquired new knowledge, and this is such a good tool that they have given us first and second cycle teachers that has helped us a lot in the classroom. » (Juana)*	*« [...] they were all very interesting topics, topics where we, perhaps because we are not specialists in the area, were impressed to see the contents and learn a little more in each one of them. Each day was very useful and the topics that the teacher developed each one of them was really interesting. » (Norma)*	No negative aspects mentioned.
T2: Pedagogy	*«For me the impression is that it has been a well-organized, well-structured process. From the first session we felt the excitement of being able to attend the following sessions, it was very interesting, and it has awakened our curiosity to be able to continue implementing some knowledge that we already had but being able to convert it into something more practical with our students. » (Jorge)*	*«I liked everything that we were taught, all the didactic methodologies implemented by the teacher and the participation of all the classmates. » (Maria Teresa)*	No negative aspects mentioned.
T3: Both	*«I've never received training in this form that is so complete, because in addition to the virtual counselling, there are the face-to-face [...] classes and the mathematical and didactic techniques and strategies as well. » (Patricia)*	*«I quite liked the fact that every day that the classes were given, they always applied concrete material. We reinforced it through, let's say, playing, through moving things. [...] One of these activities was the use of the geoboard, which I found quite interesting. » (Ana)*	No negative aspects mentioned.

Figure A7: General questions



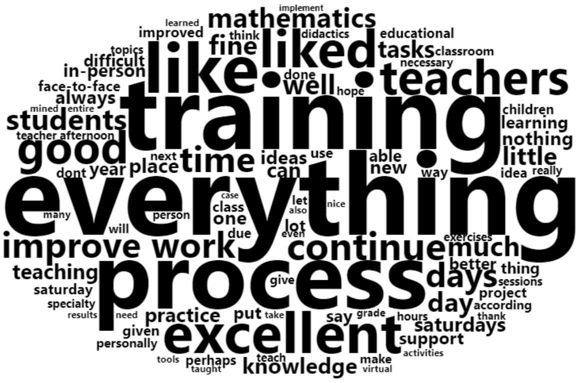
	Impact on didactical and / or mathematical knowledge	Impact on teaching	Impact on pupils	Training's impact relative to other basic education trainings
T1: Math	*«The mathematical knowledge was very nice, very broad. The trainings helped us to obtain new strategies, new techniques. [...] the training was quite playful so that we could acquire more strategic and technical knowledge for the students. » (Ingrid)*	*« [...] I have always been very persistent in making my students learn mathematics, but I didn't have the tool that would work more effectively and with this training, it has strengthened me more and I have managed to get the students to do what I wanted them to do: to like the subject. » (Juana)*	*«In this formative process with the students, they gave more importance to the subject of mathematics, and they saw it more as having it closer to them. They learnt that mathematics is beautiful, that you don't have to be afraid of it. » (Ingrid)*	*«I see it as something that helps us mathematics teachers, as we perhaps lack constant training, new strategies, and techniques to implement. » (Ingrid)*
T2: Pedagogy	*«In this process, we have been awakened to many strategies. [...] [T]hey have given us the necessary tools to be able to take them to a different level; to be able to perform better in our educational work, in our schools [...] where we work every day. » (Jorge)*	*« The most important thing, the learning pace of each student. Oftentimes we say that we do not achieve the objectives; that is where they tell us that we must look for new strategies for those [weak] students [...] because not everyone learns in the same way. » (Jesús)*	*« They really liked the games and everything, they are more motivated. [...] All this has helped us to motivate the children more and as they say, playing is also a way to learn, right? [...] it has been noticed that there is more learning. » (Jesús)*	*« We felt that that we have learned more, because of the confidence they gave us to always ask questions, the doubts we have and more than anything the tasks were well based on what we are teaching. » (Jesús)*
T3: Both	*« My speciality is not mathematics, but this process taught me a lot about mathematics [...]. For me that class was like something significant, and I won't forget how to work with fractions. » (Ruth)*	*[T]he students are no longer just in the classroom with their notebooks and the blackboard, but we also go outside the classroom [...]. For example [...] to form circles, so that they can expand their knowledge and assimilate what a perimeter is. That's already changing the old*	*«They liked it a lot, because regarding the multiplication tables game, they all wanted to go on to say the multiplication tables with the Buster game. Even the second graders felt more motivated and participated more. So, this has really changed the way we*	*« I liked that they left guides that came with many exercises that can be implemented. I took some of them up again. I feel that this made quite a difference [...] [I]t has been face-to-face, something that I have valued a lot, because virtually, you lose important*

Figure A8: Questions on Program Impact



		traditional teaching, isn't it? (Patricia)	*teach and motivate them with these dynamics. » (Patricia)*	*aspects like [...] friendships or bonds [...] with other teachers [...]. » (Johana)*
	Feedback on the seven days of face-to-face training	**Feedback on the audiovisual material provided**	**Feedback on activities between sessions**	**Feedback on the tutoring provided**
T1: Math	*«The seven days were excellent, and the explanations were very precise, and I believe that we all understood and reinforced the knowledge that they imparted to us. The teacher [...] answered our questions, and provided us with all the information we needed. » (Juana)*	*«Very applicable, practical both for us as teachers to keep practising and improving our maths skills and to be able to teach and try to make it more useful for students' understanding and teaching. » (Julissa)*	*«Some of them were a bit complicated. We needed help from materials, from the notes we made and from some of our colleagues who were always there. [...] But the truth is that everything worked well. » (Norma)*	*« It was excellent, because the teacher there was ready to explain to us each one of the topics. [...] She explained again and again and asked each one of us if we didn't understand, because [..] not all of us understand in the same way and at the same speed. » (Juana)*
T2: Pedagogy	*« Excellent, very dynamic, well-structured, everything is detailed, and the booklet helped us to be able to locate ourselves within each day. » (Jorge)*	*« Very special, because everything that was used was well-related to what we were learning, everything was very appropriate, and everything was there. » (Jesús)*	*« I think they are well structured, and they are also very short, besides the fact that one is working five days a week, sometimes there is not much time left, but in this case, they are very short and precise, and they could be solved in the medium term. » (Jorge)*	*« A well-trained person, she was well updated, and the tutorials were quite good through the zoom platform. » (Jorge)*
T3: Both	*«The teacher was [...] very innovative, very creative. I always liked the fact that she had a lot of didactic and technological resources, so that really appealed to her. As a*	*«They were quite good, because they served as a resource for those contents that were a little unclear to us in the class [...]. We would watch them again, review the content and use*	*«I consider them necessary, [...] because we also needed to consolidate the knowledge through the guides that we had. That we had to present the portfolio and the exercises, I*	*«I thought it was good, because the teacher has been very patient with us, she was very accessible and all the doubts we had, we asked her, and she explained to us with a lot of*

Figure A9: Questions on Program Activities



Word	Frequency
excellent	12
liked	11
fine	10
good	10
like	3
nice	3
seemed	3

Figure A10: Word cloud of improvement-related feedback and co-occurrences associated with the term “everything.”

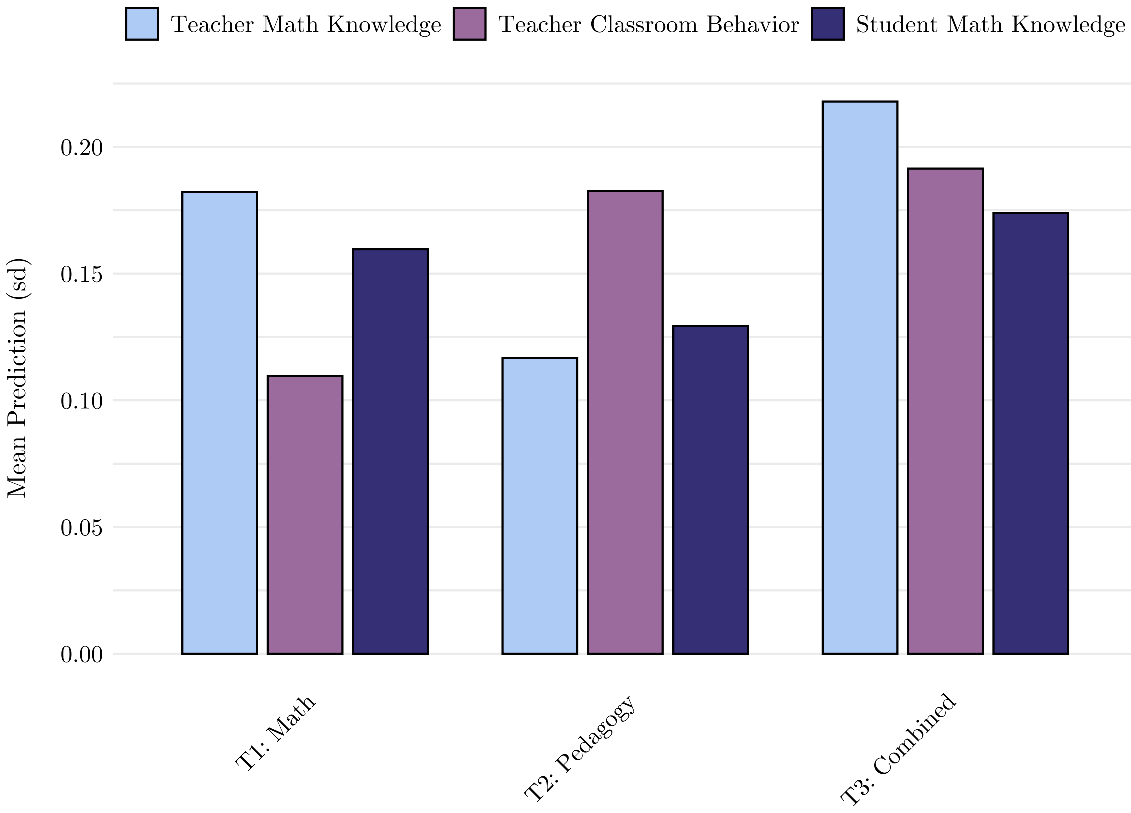


Figure A11: Predictions on program impact

Illustration based on a prediction survey with 38 education researchers and practitioners. The figure shows mean predictions of effect sizes in standard deviations per treatment. For each treatment, respondents predicted effects on teacher content knowledge, teacher classroom behavior and student content knowledge.