Jakob, Martina; Büchel, Konstantin; Steffen, Daniel; Brunetti, Aymo

Working Paper

Participatory teaching improves learning outcomes: Evidence from a field experiment in Tanzania

Discussion Papers, No. 23-10

Provided in Cooperation with:
Department of Economics, University of Bern

Suggested Citation: Jakob, Martina; Büchel, Konstantin; Steffen, Daniel; Brunetti, Aymo (2023) : Participatory teaching improves learning outcomes: Evidence from a field experiment in Tanzania, Discussion Papers, No. 23-10, University of Bern, Department of Economics, Bern

This Version is available at:
https://hdl.handle.net/10419/278646

UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

Participatory Teaching Improves Learning Outcomes: Evidence from a Field Experiment in Tanzania

Martina Jakob, Konstantin Büchel,
Daniel Steffen, Aymo Brunetti

23-10

September, 2023

DISCUSSION PAPERS

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

Participatory Teaching Improves Learning Outcomes: Evidence from a Field Experiment in Tanzania*

Martina Jakob[a,°], Konstantin Büchel[a,°],
Daniel Steffen[b], Aymo Brunetti[a]

[a]University of Bern
[b]Lucerne University of Applied Sciences and Arts

September 5, 2023

Abstract

Participatory teaching methods have been shown to be more successful than traditional rote learning in high-income countries. It is, however, less clear if they can help address the learning crisis in low- and middle-income countries, where classes tend to be large and teachers have fewer resources at their disposal. Based on a field experiment with 440 teachers from 220 schools in Tanzania, we use official standardized student examinations to assess the impact of a pedagogy-centered intervention. A five-day in-service teacher training on participatory and practice-based methods improved students' test scores 18 months later by 0.15σ. The additional provision of laptops with a learning software allowing teachers to refresh their content knowledge did not yield further learning gains for students. Complementary results from qualitative surveys and interviews suggest that the program was highly appreciated by different stakeholders, but that participants are unable to assess its impact along different dimensions, giving equally positive evaluations of its successful and its less successful elements.

JEL classification: C93, I21, J24, O15.
Keywords: productivity in education, participatory teaching, teacher content knowledge, computer-assisted learning, development economics.

*The authors acknowledge generous funding by the IMG Stiftung. This study received IRB approval from the Faculty of Business, Economics and Social Sciences at the University of Bern on November 4, 2019 (serial number: 122019). A randomized controlled trials registry entry is available at: `https://www.socialscienceregistry.org/trials/4959`.
Contact Details (°shared first authorship):
°Jakob: Univ. of Bern, Inst. of Sociology, Fabrikstr. 8, CH-3012 Bern, martina.jakob@unibe.ch
°Büchel: Univ. of Bern, Dept. of Economics, Schanzeneckstr. 1, CH-3001 Bern, konstantin.buechel@unibe.ch
Steffen: Lucerne University of Applied Sciences and Arts, Suurstoffi 1, CH-6343 Rotkreuz, dani.steffen@hslu.ch
Brunetti: Univ. of Bern, Dept. of Economics, Schanzeneckstr. 1, CH-3001 Bern, aymo.brunetti@unibe.ch

1 Introduction

Only 4 percent of students in low-income countries, compared to 95 percent in high-income countries, reach minimum literacy skills towards the end of primary school (World Bank, 2018, p. 8). To narrow the global learning gap, we need to rethink the strategies that teachers in developing countries use in the classroom. While schools in high-income countries have increasingly adopted participatory pedagogical approaches with a high degree of student engagement, more teacher-centered approaches such as lecturing and rote learning are still the norm in many low- and middle-income countries. Modern pedagogy takes a clear stance and considers student engagement a vital component of effective teaching, a view that is corroborated by vast evidence from high-income countries (e.g. Cornelius-White, 2007; Seidel and Shavelson, 2007; Harbour et al., 2015). However, it is not clear if this insight can be transferred to low- and middle-income countries, where teachers often have to manage very large classrooms and have few teaching aids at their disposal. Under such constraints, switching to more demanding teaching strategies could even prove detrimental (e.g., Berlinski and Busso, 2017). Moreover, in light of recent evidence on insufficient subject mastery among many teachers in disadvantaged regions (e.g., Sinha et al., 2016; Bold et al., 2017a; Brunetti et al., 2023), it remains an open question whether improving pedagogy alone is effective or if shortfalls in teachers' content knowledge need to be tackled simultaneously.

To address these questions, we conducted a randomized controlled trial (RCT) with 440 math teachers and more than 25,000 students from 220 schools in Tanzania. With an average of 51 students per teacher and a persistent shortage of classrooms and teaching aids, Tanzania faces resource constraints that are typical for many education systems in low-income countries (UNESCO, 2022). The intervention we study consisted of a five-day in-service program where teachers learned how to engage their students more actively in classes, bring their teaching closer to every-day live, and collaborate in teams to handle large classrooms and exchange on teaching techniques. After the initial five-day workshop, all teachers were invited to half-yearly refresher meetings to revise concepts and discuss implementation issues. Half of the teachers in the treatment group were randomly selected to further receive a laptop with a computer-assisted learning (CAL) software enabling them to refresh their content knowledge. The learning software consisted of short math videos and quizzes from "Khan Academy", and teachers participated in additional sessions to familiarize themselves with the program and discuss their progress. Both versions of the treatment were administered by Swiss NGO Helvetas that has implemented teacher training programs in Tanzania since 2000.

We scraped student-level data from standardized assessments published by National Examinations Council of Tanzania (NECTA) to estimate the impact of the program on students, and used data from our own assessments to study intermediate effects on teachers. Our design allows us to analyze direct effects on participating teachers and their students as well as spillover effects on peer teachers and students in treated schools. To better understand the mechanisms behind potential effects, we complemented the experimental data with classroom observations, surveys, and in-depth interviews.

Our analysis establishes four sets of findings: First, switching to participatory pedagogy successfully improved overall student tests scores two years later by 0.15σ (p-value=0.018), and the share of students with top grades increased by 6 percentage points from 16 to 22 percent (p-value=0.013). Point estimates for pass rates are positive too, but do not reach statistical significance (p-value=0.117). These effects are particularly remarkable considering that we used data from official national tests

that were not specifically tailored to the intervention. Our complementary data shows that treatment teachers did indeed apply a wide range of the participatory pedagogical strategies taught in the training, such as group work (observed in 87% of classroom visits), games (28%) or dialogue (26%), and expressed great enthusiasm for the program in in-depth interviews.

Second, students who were taught by teachers equipped with laptops and CAL software did not outperform students whose teachers only participated in the pedagogical intervention. Point estimates for the difference between the teacher in-service training with and without supplying the CAL software are small and statistically insignificant. While teachers receiving the laptop with CAL software markedly improved their understanding of concepts related to the subdomain of number sense and arithmetic by 0.22σ (p-value=0.058), the effect on an overall score of math proficiency is statistically insignificant (p-value=0.135). The average teacher achieved 78 percent correct answers at baseline, suggesting that many teachers were already sufficiently proficient in their subject before the intervention.[1] This is in line with results from our heterogeneity analysis showing that the CAL based refresher was significantly more effective for teachers with low content knowledge at baseline.

Third, we do not find evidence for spillovers on indirectly exposed teachers and students in treatment schools, even though the program was specifically designed to produce such externalities. Although trained teachers and their peers self-reported that they engaged in cascading activities such as model lessons and peer learning groups, estimates for spillover effects at both the student and the teacher level are close to zero and statistically insignificant (p-value=0.403).

Fourth, the data from our complementary analyses allows us to compare participants' views about impacts of the program with the actual causal estimates from the RCT. We observe that participants' survey and interview responses are not very informative about what aspects of the program did or did not work, as respondents gave equally positive evaluations for all of them. For example, while 74 percent of the trained teachers strongly agree with the statement that the program improved their pupils' math skills, so do 78 percent of their indirectly exposed colleagues, even though we do not find any indication for such spillovers in our experimental data.

Our study contributes to a growing body of literature on how to address the learning crisis in developing countries. A vast spectrum of approaches has been evaluated in recent decades (see, e.g., Kremer et al., 2013; Glewwe and Muralidharan, 2016; World Bank, 2018, for an overview), but one key factor has received surprisingly little attention: teachers. Closing the global learning gap will crucially depend on how teachers in low- and middle-income countries perform in the classroom. The pivotal role of teachers in developing countries has been appreciated by recent studies focusing on the role of teacher incentives and pay, including De Ree et al. (2018), Duflo et al. (2012), Mbiti et al. (2019b), and Muralidharan and Sundararaman (2011). Yet, the teacher performance not only depends on the economic incentives instructors face, but also on the repertoire of teaching strategies they have at their disposal.

A common strategy pursued by many development agencies is the promotion of a more student-centered pedagogy. Our study provides support for this approach, suggesting that attending five days of training in participatory pedagogy can be enough for teachers to restructure their classes and achieve higher learning gains for their students – even when their classes are large and few teaching aids are readily available. Promoting more engaging teaching strategies in low- and middle-income

[1] It is noteworthy that this substantially higher than the performance of teachers in El Salvador who averaged 47 percent on an almost identical assessment (Brunetti et al., 2020).

countries may thus be an essential element in the global quest for "inclusive and equitable quality education" (UN, 2015).

Our paper also ties into a nascent strand of literature studying complementarities in the educational production function (e.g., Mbiti et al., 2019a). Our findings suggest that shortfalls in teacher content knowledge are unlikely to constitute a binding constraint to effective teaching in Tanzanian primary schools. Teachers already exhibited considerable subject mastery, and the pedagogy intervention was at least equally successful in improving student learning without simultaneously addressing shortfalls in content knowledge.

We also add to the literature on treatment externalities. The canonical example for treatment externalities in education was documented by Miguel and Kremer (2004), where treating students with de-worming pills produced large spillovers on non-targeted children such as younger siblings. Such treatment externalities can drastically boost the cost-effectiveness of an educational program, a fact that has given rise to so called *cascading models* to deliberately include the promotion of spillovers in program designs. Our findings suggest that in the context of pedagogical interventions, achieving such externalities may not be straightforward. A possible explanation is that teachers need a considerable degree of (first-hand) exposure to the new teaching strategies to be able and willing to effectively restructure their classes.

Finally, this paper contributes on the methodological discussion on how best to evaluate programs (Banerjee and Duflo, 2009; Garbarino and Holland, 2009). While qualitative methods such as surveys and interviews provide important insights and fruitfully complement experimental data, our findings suggest that they may be ill-equipped to assess the impact of a program and distinguish between its successful and less successful elements. This highlights the importance of quantitative analysis to learn what actually works rather than relying on people's self-reports about it.

2 Context and Intervention

Our study is set in Tanzania, a lower-middle income country in East Africa. Tanzania's education system faces several challenges that are typical for developing countries. The massive expansion of schooling starting in the late nineties has put considerable strain on schools throughout the country, and resulted in shortages of teachers, classrooms and teaching materials. Consequently, the pupil-teacher ratio in primary schools stands at 51 students per instructor (UNESCO, 2022). In this context, the country has struggled to translate enrollment into learning. For example, about sixty percent of students in grade 3 are unable to read and understand a simple paragraph (Sumra et al., 2015). Learning outcomes crucially depend on what teachers do in the classroom. However, a recent study finds that only 36 percent of teachers in Tanzania possess the minimum pedagogical knowledge needed for effective teaching (Bold et al., 2017a).

The program we study in this paper was implemented by Helvetas, a large Swiss development organization focusing on building capacity in Africa, Asia, Latin America and Eastern Europe. Helvetas has been active in Tanzania for more than 50 years with projects in a broad range of fields including agriculture, youth employment, and education. After several years of piloting teacher professional development at small scale, Helvetas, the Tanzanian Teachers' Union (TTU), and the Ministry of Education jointly launched the SITT program (Inclusive School-Based In-Service Teachers Training) aiming at transforming pedagogy in Tanzanian classrooms. Prior to the experimental evaluation we

discuss in this paper, the program had already been rolled out in 1,430 schools throughout North-eastern Tanzania.

The aim of the program is to promote a more *student-centered approach* to teaching that fosters active participation among pupils. This involves activities such as group work or students taking turns with the teacher to explain concepts in front of the class. To make classes more accessible and relevant to students, teachers are encouraged to incorporate practical examples from everyday life. Through the use of inexpensive local materials such as berries, stones or toothpicks, teachers also learn how to address shortages in high-quality teaching aids. These strategies are conveyed to teachers and to the responsible government officials through a centrally organized five-day workshop. After the initial training, teachers are invited to participate in biannual two-day refresher meetings, where the application of the strategies is discussed and experiences are shared. As a guide throughout the school year, each teacher receives a comprehensive manual summarizing the teaching strategies.

In the spirit of a *cascading model*, participating teachers are also encouraged to share their knowledge with all other teachers in their schools through different collaborative activities. Most importantly, they are expected to invite their colleagues to model lessons to showcase the new teaching methods in action. Trained teachers also have to organize peer learning groups where their peers can discuss their impressions from the model lessons and share their experience with the new pedagogical techniques in their own teaching. Finally, teachers are encouraged to manage large classes as a team to promote cooperative behavior and joint learning. The implementation of the new teaching strategies and the cascading activities is overseen by government quality assurance officers and the Helvetas team through monitoring visits to targeted schools. As an indirect monitoring tool, teachers are added to a "WhatsApp" group where they are expected to share their experiences.

In 2020, the intervention was supplemented by additional activities to address potential shortfalls in teachers' content knowledge. In this context, half the teachers participated in an extended version of the program where they received a laptop equipped with a *computer-assisted learning* software. Learning materials included video content and short quizzes in Swahili produced by Khan Academy and were provided through the offline-first learning platform Kolibri developed by Learning Equality. Learning videos were typically around 5 to 10 minutes long and structured into three broad themes, *(i)* Number Sense and Elementary Arithmetics (NSEA, 80 videos), *(ii)* Geometry and Measurement (GEOM, 80 videos), and *(iii)* Data, Statistics and Probability (DSP, 11 videos). Videos were shared through a user-friendly interface and complemented with short quizzes. Each quiz drew on a basis of roughly 20 items that were presented in random order. Upon submitting an answer, users received instant feedback. The software tracked performance and awarded badges of success for quizzes with at least five correct answers. Previous studies have shown computer-assisted studying with Khan Academy to be effective at improving test scores of both students Büchel et al. (2022) and teachers (Brunetti et al., 2023).

3 Research Design

3.1 Sampling and Randomization

To assess the impact of the in-service teacher training, we conducted a randomized controlled trial with a sample of 220 public primary schools in the Tanzanian districts in of Mbulu DC, Mbulu TC,

Karatu and Siha, where the program had not been introduced yet. The implementing organization adopted a selection protocol similar to earlier implementation phases by excluding the best performing and the geographically least accessible schools in each district.

The experimental design allows to distinguish between *direct effects* on participating teachers and their pupils as well as *cascading effects* on peer teachers and their pupils. Specifically, selected schools nominated two teachers for the study: one *targeted teacher* for possible program participation and one *peer teacher* who was included for the estimation of spillovers. The selection of both targeted and peer teachers was done in coordination with the district education office and tied to the conditions *(i)* that both teachers should instruct math, and that *(ii)* the *targeted teacher* should teach math to sixth grade pupils in 2020 and seventh grade pupils in 2021. This procedure yielded a total sample of 440 teachers from 220 schools.

After the selection of schools and teachers, the research team randomly assigned each of the 220 schools to one out of three experimental conditions (see Figure 1):

- Pedagogy (65 schools, 130 teachers): Targeted teachers participated in the pedagogy training and were instructed to share their knowledge with their colleagues at their school.
- Pedagogy + Content (65 schools, 130 teachers): Targeted teachers participated in the pedagogy training and were instructed to share their knowledge with their colleagues at their school. They also obtained a laptop with computer-assisted learning software to self-study math.
- Control (90 schools, 180 teachers): Targeted teachers did not participate in any intervention activities.

Randomization was conducted after the nomination of teachers and the baseline data collection, and was stratified along three dimensions: district of school, baseline performance of pupils (i.e., school average in the standard 4 national examinations in 2018), and baseline performance of targeted teachers (i.e., math assessment conducted in November 2019).

3.2 Data

We rely on nationally standardized tests to measure effects on students, and conducted our own assessments to study intermediate effects on teachers. This experimental data is complemented with qualitative data we collected through classroom observations, surveys, and interviews in the treatment group.

Student assessments. The National Examinations Council of Tanzania (NECTA) conducts two standardized national student assessments that can be leveraged for this study: the *Primary School Leaving Examination* (PSLE) administered in grade 7, and the *Standard Four National Assessment* (SFNA) administered in grade 4. These yearly assessments are conducted with the entire student population in the respective grades and have high stakes: failing SFNA requires pupils to repeat grades, and passing PSLE is mandatory for admission in secondary school. Both assessments cover various subjects, but we rely on math scores for the main analysis. The math module in PSLE consists of 45 items that need to be completed in two hours, and SFNA includes 25 math questions students

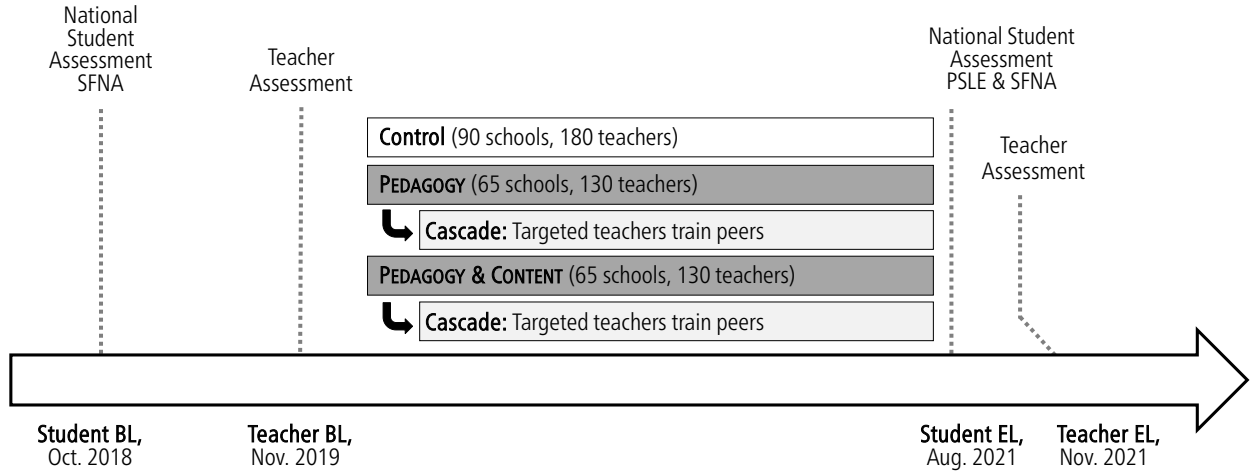


Figure 1: Timeline of the study.
The main intervention event is a five day workshop for all treated teachers and was conducted in February 2020. Afterwards, teachers implemented the new strategies and share them with their colleagues, participated in biannual meetings, and were visited by quality assurance officers of the Ministry of Education. The *National Standard Four Assessment* (SFNA 2018, SFNA 2021) and the *Primary Standard Leaving Examination* (PSLE 2021) are conducted by the Tanzanian government and the results are published online, see `https://onlinesys.necta.go.tz/`.
Source: Own representation.

need to answer in 90 minutes (NECTA, 2018, 2020). Assessment data is publicly available at the student level.

Our *main outcome measure* is the PSLE math score of seventh graders in 2021, the cohort taught by targeted (and potentially trained) teachers in 2020 and 2021. Pupils' PSLE scores can be merged with their SFNA scores from three years earlier (i.e., 2018) to establish a pupil-level baseline score. To assess spillover effects through *cascading*, the SFNA math scores from grade four pupils in 2021 can be used, as these pupils were taught by peer teachers in the same school who were exposed to cascading activities.[2] No baseline data is available in this case. As both PSLE and SFNA results are published online, we use web scraping to obtain the student-level data. Our final sample consists of 10,101 seventh graders to assess the direct effects of the programs and 15,023 fourth graders to estimate spillovers.

Teacher assessments. To measure teacher content knowledge in math, all 440 study participants were invited to two comprehensive math assessments conducted before and after program implementation. The assessments were designed to mirror the Tanzanian primary school curriculum between grade 2 and grade 7 and covered the domains Number Sense & Elementary Arithmetics (NSEA, about 60%), Geometry & Measurement (GEOM, about 35%), and Data, Statistics, & Probability (DSP, about 5%). Assessments were administered as paper-and-pencil tests in regional meet-ups and had to be completed in 90 minutes.

[2]Note that standard 4 pupils were not necessarily taught by the one peer teacher who was chosen to participate in the teacher assessments. However, this is irrelevant for the study of pupils' learning outcomes as cascading activities are explicitly targeted at all teachers in a school and hence should impact learning across all grades and classrooms in a program school.

Complementary qualitative data. We collected three different types of qualitative data to get deeper insight into how switching to participatory pedagogy was viewed and put in practice by treated teachers. First, all teachers had to fill in a short survey about their evaluations of the program and their perceptions about how it had impacted them and their students. The survey primarily included single-choice questions, where respondents could rate certain elements or indicate whether they agreed or disagreed with a given statement, but also featured space for written feedback and suggestions. Survey forms were administered during the endline math assessment to all teachers and tailored to the different experimental groups.[3] Second, to better understand how teachers incorporated the new methods into their classes, quality assurance officers of the education ministry conducted classroom observations in lessons of program participants. Based on the TEACH tool proposed by the World Bank (2019), a monitoring questionnaire was designed and government officials were briefed on how to conduct the classroom observations. Overall, 112 visits to treated teachers were conducted. Third, to complement the surveys and interviews, six participants of the PEDAGOGY intervention (about 120 min. audio recordings), six teachers from the PEDAGOGY & CONTENT group (about 120 min. audio recordings), six peer teachers (about 70 min. audio recordings), and twelve government or TTU officials (about 150 min. audio recordings) participated in *semi-structured interviews.*[4]

3.3 Baseline characteristics, compliance, and attrition

Table A.1 in the appendix shows that *baseline characteristics* are well-balanced across the three experimental groups. The average teacher in our sample scored 78 percent correct answers on the math test we administered prior to the intervention. As the test was designed to cover the Tanzanian primary school curriculum, this suggests that, on average, teachers master three quarters of the materials they have to teach. About 4 in every 10 teachers in our sample are female and the average teacher is 38 years old. Panel 2 on school characteristics shows that the typical class size is about 40 students.[5] The number of students that took the SFNA exam, roughly 50 per school, provides a proxy for the number of students per grade. As this figure is not much higher than the average class size, most schools can be assumed to have only one class per grade. Most importantly, pupils' baseline scores are well-balanced across experimental groups. On average, about 67 percent of students passed the baseline math exam, and 40 percent of students scored one of the two top grades (A or B).

Our monitoring data suggests that *compliance* with the treatment assignment was very high. All

[3]We designed four different questionnaires: (1) a questionnaire for teachers in the PEDAGOGY treatment with items about the training and the implementation of the new methods, (2) a similar questionnaire for the PEDAGOGY + CONTENT group with additional questions about the content training with the laptops, (3) a questionnaire for peer teachers asking about casacading activities, and (4) a short questionnaire for the control group with questions about the evaluation process. With the exception of the control group, the different survey versions followed the same basic structure and had many common items, allowing for comparison across different groups.

[4]During these conversations, the interviewees were asked *(i)* to share their general impression of the intervention, *(ii)* to explain their view on the main elements of the PEDAGOGY intervention, *(iii)* to share their assessment on the impact of the program on teachers' math and teaching skills as well as the learning outcomes of children, and *(iv)* to give feedback on selected activities and program inputs; additionally, officials were asked *(v)* to compare the pedagogical intervention with similar educational initiatives by other organizations, and *(vi)* to comment on their attitude towards rigorous program evaluation. Table D.1 in the appendix section D provides an overview of statements by topic and type of interviewee.

[5]While information on the number of pupils per classroom is difficult to collect, the number of pupils per *stream* can serve as a proxy. In Tanzania the concept of a "class" is surprisingly blurry because several streams of pupils can be instructed in one classroom (and effectively become one class) if schools do not have enough classrooms or teachers to teach streams separately.

teachers in the treatment group participated in the five-day teacher training, and 94 percent of the teachers in the PEDAGOGY & CONTENT group report having used the laptops for content revision. To be able to assess the impact of the program using students' tests scores in grade 7, targeted teachers had to teach math to all sixth graders in their school in 2020 and to all seventh graders in 2021. Our data collected during the endline teacher survey shows that 85 percent of the students in the treatment group were indeed taught by targeted teachers. This share does not differ significantly between experimental groups.

Tables A.2 and A.3 examine patterns of *attrition* for teachers and students respectively. At the teacher level, 99 percent of the selected teachers took part in the baseline assessment, and attrition for the endline assessment was about 15 percent and evenly distributed across experimental groups. This yields a total sample size of 368 teachers. At the student level, we start with baseline data for 12,657 pupils from 220 schools. About 17 percent of these students either dropped out of school between grade 4 and grade 7, missed the endline examination, or could not be matched between the two examination rounds. Moreover, one school dropped out because the targeted teacher missed both the base- and endline data collection. Finally, an estimation sample with 10,101 seventh graders from 219 schools remains. Both for teachers and pupils, attrition was unrelated to the experimental assignment. For the estimation of spillovers, we can use a sample of 15,023 grade 4 students from 220 schools. Due to the unavailability of baseline data, we cannot study the attrition for this cohort of students.

4 Results

4.1 Did promoting participatory teaching strategies improve learning?

We estimate the *intent to treat* (ITT) effect on students of directly targeted teachers with the following benchmark equation

$$Y_{isk}^{PSLE} = \beta Treatment_s + X_i^{'}\gamma + V_s^{'}\lambda + \phi_k + \epsilon_{isk}, \tag{1}$$

where Y_{isk}^{PSLE} is the standardized math PSLE score of student i in school s and stratum k at endline, and *Treatment* is a binary indicator that takes the value of 1 if a school was assigned to the treatment group and is 0 otherwise. Student level controls, X_i, comprise sex, baseline math score, and average baseline score across all subjects taken from the SFNA baseline assessment. V_s represents a vector of school-level controls including the number of students who took the baseline assessment, the average PSLE score at baseline[6], the driving distance to the district headquarters and the class size, as well as the math score, sex, and age of the targeted teacher. ϕ_k stands for k strata fixed effects, and ϵ_{isk} represents the error term.

The results in Table 1 document that students in treated schools significantly outperformed the control group by 0.15σ (column 2). Pupils in program schools were also up to 6 percentage points more likely to achieve a top grade (i.e., A or B) than their peers in control schools (columns 3 and 4). This corresponds to an increase in top grades by 36 percent. Estimates in columns 5 and 6 further

[6]Note that this is not the average score of the cohort we study, but that of a previous cohort of seventh graders in the school.

Table 1: Overall program effect on the math score of pupils

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.107^{+}	0.145*	0.046*	0.056*	0.023	0.036
	(0.062)	(0.061)	(0.023)	(0.022)	(0.023)	(0.023)
Pupil baseline math score	0.466**	0.327**	0.121**	0.082**	0.210**	0.155**
	(0.017)	(0.021)	(0.008)	(0.008)	(0.008)	(0.010)
Mean of dep. variable	-0.008	-0.008	0.155	0.155	0.592	0.592
Observations	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.252	0.295	0.146	0.180	0.202	0.224
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized math score for columns (1) and (2), a binary variable indicating whether a pupil scored A or B (highest grades) in math for columns (3) and (4), and a binary variable indicating whether a pupil passed the math exam for columns (5) and (6). *Pupil baseline math score* is a pupil's score in the SFNA exam administered in grade 4. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils in grade 4 and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

suggest that the program induced a 2 to 4 percentage point increase in pass rates, but these effects are not statistically significant at conventional levels.

Table A.5 in the appendix examines effects on students' average score across all subjects rather than their math score. Results are very similar, with estimated effects of 0.12σ and an increase in top grades by 7 percentage points or 30 percent. This suggest that although the pedagogical training was tailored to math, teachers were able to transfer the methods to other subjects.

Overall, the observed impacts are comparable to effects documented in RCTs of similar programs (see Snilstveit et al., 2015; McEwan, 2015). Unlike most other studies, our analyses are based on standardized national assessments that are not tailored to the intervention under study, which strengthens their external validity.

Our causal estimates are consistent with insights from our complementary data sources. Classroom observations point to a widespread use of the participatory teaching strategies advertised through the training program. As Figure C.1 in the appendix shows, treated teachers frequently applied methods such as group work (87% of visits), games (28%), student presentations (28%), and dialogues (26%). Treatment teachers also used a wide range of teaching materials, including daily life objects (66% of visits), textbooks (46%), and flash cards (20%). The survey data further shows that 96 percent of treated teachers rate the participatory teaching model as excellent (75%) or good (21%). Similarly, 96 percent of targeted teachers strongly (74%) or rather agree (22%) with the statement that the intervention improved their students' math scores. The high appreciation for the program also surfaced in the interviews where teachers often used words such as *"improve"*, *"change"*, and *"enjoy"* when talking about the intervention (see Table D.1 in the appendix).

To better understand under which circumstances the participatory teaching methods promoted through the training work best, it is informative to take a look at how effects vary by characteristics

Table 2: Program effect on the math score of pupils by implementation version

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Pedagogy	0.127	0.147*	0.056+	0.059*	0.024	0.033
	(0.081)	(0.071)	(0.029)	(0.026)	(0.028)	(0.026)
T2: Pedagogy & Content	0.086	0.142+	0.034	0.052+	0.022	0.039
	(0.072)	(0.073)	(0.026)	(0.027)	(0.029)	(0.028)
Pupil baseline math score	0.466**	0.327**	0.121**	0.082**	0.210**	0.155**
	(0.017)	(0.021)	(0.008)	(0.008)	(0.008)	(0.010)
T2 − T1	-0.041	-0.005	-0.022	-0.008	-0.002	0.006
	(0.090)	(0.075)	(0.033)	(0.028)	(0.033)	(0.030)
Mean of dep. variable	-0.008	-0.008	0.155	0.155	0.592	0.592
Observations	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.252	0.295	0.147	0.180	0.201	0.224
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized math score for columns (1) and (2), a binary variable indicating whether a pupil scored A or B (highest grades) in math for columns (3) and (4), and a binary variable indicating whether a pupil passed the math exam for columns (5) and (6). *Pupil baseline math score* is a pupil's score in the SFNA exam administered in grade 4. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils in grade 4, and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

of classes, teachers and pupils. A key challenge for productive student engagement is posed by the typically very large classes in Tanzania. According to Table A.7, the impact of the interventions decreased with larger class sizes, but these effects are not statistically significant (columns 7 and 8). A further concern might be that the use of participatory teaching methods demands a high level of skills on the part of the teachers. We do not observe teachers' pedagogical skills, but their performance in the math test can serve as a proxy. Indeed, treatment effects appear to be larger for students who are taught by better-performing teachers (columns 5 and 6). Additional analyses by pupils' gender and initial performance levels do not point towards relevant effect heterogeneity along these dimensions.

4.2 Did the computer-based content training yield additional benefits?

We also estimate the effects of each program version separately, using

$$Y_{isk}^{PSLE} = \beta_1 T1_s + \beta_2 T2_s + X_i^{'}\gamma + V_s^{'}\lambda + \phi_k + \epsilon_{isk}, \tag{2}$$

where $T1_s$ is a binary indicator for the Pedagogy intervention, and $T2_s$ indicates whether a treated teacher's school was additionally assigned to the content training component, i.e. to Pedagogy & Content.

As Table 2 shows, we do not find that providing laptops for content revision in addition to the pedagogical training yielded further learning gains for students. If anything, the point estimate for the extended intervention is slightly lower, but this difference is not significant.

One possible interpretation is that teachers did not use or appreciate the laptops for the intended purpose. Our complementary data suggests otherwise. Teachers report spending an average of 5 to 6 hours per week with the learning software, and provide very positive evaluations of the computer-assisted learning component with 68 percent rating it as excellent and 20 percent as good. The same affirmative feedback surfaced in interviews, where teachers unanimously expressed strong appreciation for the laptops and reported using them frequently for content revision or to prepare their lessons.

Another possibility is that teachers did use the laptops, but failed to meaningfully improve their content knowledge with the software. Figure 2 and Table A.8 present estimates for the causal impact of each intervention on teachers' content knowledge in math. Although teachers in the laptop group markedly improved their understanding of concepts related to NSEA by 0.22σ (columns 5 and 6 in Table A.8), the effect on an overall score of math proficiency is smaller (0.15σ) and misses conventional levels of statistical significance (columns 1 and 2).

A plausible interpretation for these modest effects is that most teachers already possessed good mastery of the primary school curriculum to begin with. As indicated in Figure A.2, the average teacher was able to answer 78 percent of the questions on materials covered in grades 2 to 7 correctly. While targeted teachers scored an average of 81 percent, peer teachers scored only 74 percent, suggesting that schools selected particularly well-performing teachers for program participation. Overall, 50 percent of the teachers pass the threshold for subject proficiency – at least 80 percent correct answers – advocated by the World Bank (Bold et al., 2017a). Only 2 percent of all teachers answered less than 50 percent of the questions correctly. A comparison with results from an almost identical assessment conducted with teachers in El Salvador suggests that the Tanzanian teachers perform considerably better than their counterparts in El Salvador (see Brunetti et al., 2020).[7] Hence, it appears plausible that many Tanzanian teachers are already sufficiently proficient in math for effective teaching at the primary school level. In line with this argument, Table A.9 in the appendix points to considerable effect heterogeneity by teachers' initial ability level. Low-performing teachers markedly improved their content knowledge (0.51σ, p = 0.004, for teachers below the median) due to the intervention, but these effects decline significantly as teachers' baseline scores improve, and are close to zero for high-performing teachers (not shown).

Hence, from an impact evaluation perspective, the additional investment in the IT equipment for content revision clearly did not pay off. Although we provide suggestive evidence that low-performing teachers used the software to catch up with their better-prepared colleagues, we do not find that such gains were transferred to students.

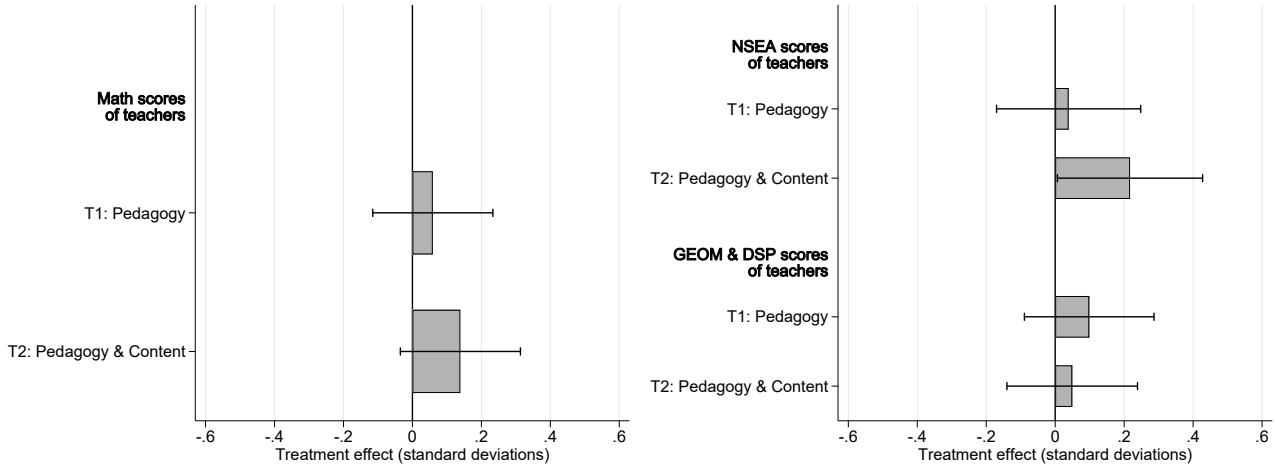
4.3 Did the interventions produce externalities for indirectly exposed students and teachers?

To estimate spillovers on indirectly exposed fourth-graders rather than directly exposed seventh graders, we use the following slightly adapted version of equation (1)

$$Y_{isk}^{SFNA} = \beta Treatment_s + X_i^{'}\gamma + V_s^{'}\lambda + \phi_k + \epsilon_{isk}, \tag{3}$$

where Y_{isk}^{SFNA} is the standardized math SFNA score of student i in school s and stratum k at

[7] The average teacher in the El Salvador study scored 47 percent on a math test covering materials from grades 2–6 and only 14 percent of teachers achieved at least 80 percent correct answers.



(a) Effects on overall test score

(b) Effect on test scores by domain

Figure 2: Treatment effects on teachers' overall and domain-specific math scores
Estimates for the effect of the two intervention versions on *targeted teachers* are shown. *Controls* include baseline score, sex, age, and years since graduation at baseline. 90 percent confidence intervals shown. For more information on the sample size and the estimation strategy, see Table A.8.

endline. As no nationally standardized assessment results are published for students below grade four, we include the school-level SFNA score as a baseline performance measure.

Table 3 examines spillover effects on students whose teachers were indirectly exposed to the treatment through peer learning activities in their school. In all specifications, estimates are close to zero and insignificant. In line with the moderate direct effects of the additional content training, we also find no indication of content knowledge spillovers at the teacher level, as Table A.8 shows.

A possible explanation for the absence of meaningful treatment externalities is that the observation period of our study was not long enough to capture effects on students of indirectly exposed teachers. Due to the time lag between the initial teacher training and the cascading activities, peer teachers may not have had sufficient time to put the new techniques into practice. To assess the plausibility of this hypothesis, we can draw on non-experimental data from the implementation phase 2013 to 2019, i.e. the period prior to the execution of the field experiment. Using both the PSLE and the SFNA scores for these years, we conduct a difference-in-difference analysis to assess the impact of the program over a longer time horizon (see Appendix B). As only one out of many teachers in each intervention school participated in the teacher training and all other teachers were indirectly exposed through cascading activities, our estimates correspond to an upper bound for spillover effects at the school level. As Table B.1 in the appendix shows, we find no indication for such effects.

Another possibility is that the knowledge sharing activities were not conducted. Again, our complementary data suggests otherwise. Almost all targeted teachers report organizing the model lessons (95%) and the peer learning groups (96%), and most peer teachers report participating in these activities (88% for both model lessons and peer learning groups), with the average peer teacher claiming to have attended 3.8 model lessons. Moreover, the knowledge sharing activities are rated very positively by both targeted and peer teachers.[8]

[8]This should not be seen as conclusive evidence for the successful implementation of the cascading elements as teachers may have succumbed to a common tendency of giving socially desirable, but dishonest answers. Indeed, in the in-depth interviews, teachers provided slightly more critical feedback on the cascading elements, with some interviewees mentioning challenges regarding their implementation due to the lack of interest of some of their colleagues.

Table 3: Cascading effect on the math score of pupils

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.028	0.037	0.011	0.012	0.011	0.016
	(0.048)	(0.044)	(0.009)	(0.009)	(0.021)	(0.019)
School PSLE avg. score (std)	0.081*	0.083**	0.016**	0.018**	0.034**	0.033**
	(0.031)	(0.031)	(0.006)	(0.006)	(0.013)	(0.012)
School SFNA avg. score (std)	0.134**	0.129**	0.022**	0.022**	0.048**	0.046**
	(0.031)	(0.031)	(0.006)	(0.006)	(0.013)	(0.013)
Mean of dep. variable	-0.000	-0.000	0.075	0.075	0.368	0.368
Observations	15023	15023	15023	15023	15023	15023
Adjusted R^2	0.072	0.080	0.035	0.040	0.053	0.060
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized SFNA math score for columns (1) and (2), a binary variable indicating whether a pupil scored A or B (highest grades) in math for columns (3) and (4), and a binary variable indicating whether a pupil passed the math exam for columns (5) and (6). *School-level baseline scores* are the school's average scores in the SFNA exam administered in grade 4 and the PSLE exam administered in grade 7. Controls include *(i) pupil-level controls* for sex, *(ii) school-level controls* for the number of pupils in grade 4 and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

Hence, a more likely explanation is that although the cascading activities were conducted, they did not provide sufficient exposure to the new pedagogical techniques for peer teachers to effectively restructure their classes.

4.4 How informative are participants' self-reports about the impact of different program aspects?

An ongoing debate in the development community concerns the merits of two distinct evaluation traditions: a quantitative paradigm emphasizing causal inference methods and a qualitative tradition focusing on the experiences of project stakeholders (e.g., Banerjee and Duflo, 2009; Garbarino and Holland, 2009). The main contribution of this paper is quantitative, but we can also combine and compare our experimental findings with insights from qualitative surveys and interviews with project beneficiaries. In particular, we asked all participating teachers to assess the effect of the intervention on different outcomes, allowing us to contrast these self-reports with the actual causal effects we identified through the experiment (see Table 4).

Across all the outcomes and groups we study, participants are very confident about the impact of the intervention. While this is in line with the positive causal impact we report, response patterns appear to be unrelated to the success and failure of different project components. Most notably, directly participating teachers and peer teachers are equally optimistic about the impact of the intervention on their math skills and those of their students, even though we find no indication for spillover effects in our data. Similarly, we report no effect of the Pedagogy intervention on teachers' math skills, but 87 percent of teachers in this group strongly agree with the claim that they improved these skills

Table 4: Comparison between observed causal effects and participants' reported beliefs

	RCT: Observed impact	**Survey: Participants' beliefs about impact**
Impact of intervention on student learning	Significant effect of 0.15 SD*	Did the project improve the math skills of your pupils? Strongly agree: 74%, rather agree: 22%
Spillovers of intervention on students of peer teachers	Effect insignificant and close to zero	Did the project improve the math skills of your pupils? Strongly agree: 78%, rather agree: 19%
Impact of Pedagogy *intervention on teachers' math skills*	Effect insignificant and close to zero	Did the project improve your math skills? Strongly agree: 87%, rather agree: 5%
Impact of Pedagogy & Content *intervention on teachers' math skills*	Effect of 0.15 SD, but insignificant	Did the project improve your math skills? Strongly agree: 85%, rather agree: 11%
Spillovers of intervention on peer teachers' math skills	Effect insignificant and close to zero	Did the project improve your math skills? Strongly agree: 81%, rather agree: 15%

thanks to the intervention. Finally, teachers rated the self-studying with the laptops very positively, but we find only limited evidence for its effects at the teacher level and no evidence for an impact on students.

These findings tie into a nascent literature studying biases in evaluations (e.g., Camfield et al., 2014). Two broad explanations accounting for participants' overoptimistic impact assessments can be distinguished. First, people's capacity for counterfactual thinking is limited, leading them to misattribute outcomes or changes in their lives to the programs they participated in (e.g., McKenzie, 2018). Comparing actual and self-reported effects in three labor market interventions, Smith et al. (2021) conclude that participants act as "lay scientists". Their assessments are largely unrelated to the actual causal impact estimated for their group, but tend to follow coarse heuristics for this impact such as unconditional outcomes or before-after comparisons. A second well-documented bias in social science research, known as courtesy bias, social desirability bias or experimenter demand effects, is a general tendency of subjects to provide answers they perceive as aligning with the researcher's expectations (Camfield et al., 2014; Krumpal, 2013; Zizzo, 2010). In project evaluation, the resulting pro-project bias is likely to be exacerbated if people believe that the evaluation will determine whether the project is continued. Our findings are in line with these biases and suggest that while qualitative evidence from participant surveys and interviews can provide a valuable complement to experimental evidence, it is ill-equipped for the assessment of causal impacts.

5 Conclusion

Addressing the learning global learning crisis calls for innovative strategies to track and improve education (e.g. Patrinos and Angrist, 2018; World Bank, 2018; Jakob and Heinrich, 2023). In this paper we turn our attention to the teachers, who are the key actors in the educational system. While previous research has strongly focused on the misaligned economic incentives teachers often face, this study is premised on the assumption that they could be using ineffective pedagogy. Through a randomized controlled trial with 440 teachers and about 25,000 students in Tanzania, we show that promoting participatory teaching strategies significantly improves students' learning outcomes by 0.15σ. Our findings are based on standardized national assessments conducted by the National Examinations Council of Tanzania and corroborated by evidence from our classroom observations and participant surveys affirming that teachers indeed implemented and appreciated the new participatory methods.

Our study also explores the potential of computer-assisted learning to improve teachers' content knowledge and, thereby, student learning. We find suggestive evidence that providing computers with a learning software helps low-performing teachers improve their math skills. However, this does not translate into measurable learning gains for their students. Previous research suggests that a 0.1σ gain student learning would require a 1σ improvement in teachers' content knowledge (Bau and Das, 2020; Metzler and Woessmann, 2012) – an unrealistically large effect for educational interventions. Our findings underscore that addressing shortfalls in teachers' content knowledge is not a low-hanging fruit for promoting student learning.

We report similarly discouraging results for spillovers on other teachers and their students through cascading activities. Cascading schemes are favored in the development community for their potential to increase the number of beneficiaries and extend a project's reach. However, our results suggest that producing measurable learning spillovers is not straightforward. More research is thus needed to explore if and how the promise of cascading can be realized in educational initiatives.

Nevertheless, even without relying on spillovers, building teacher competencies can be a very cost-effective approach to improve student learning in the long run. Teachers often remain in their profession for many years, influencing dozens of student generations. If they continue to apply the new teaching methods throughout their professional lives, pedagogical teacher training becomes a highly sustainable and cost-effective means to foster student learning. Hence, promoting participatory teaching could be a key ingredient to a comprehensive strategy to ensure that children in developing countries are not only going to school, but are actually learning.

References

Banerjee, Abhijit V and Esther Duflo. 2009. The experimental approach to development economics. *Annu. Rev. Econ.* 1 (1):151–178.

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

Berlinski, Samuel and Matias Busso. 2017. Challenges in educational reform: An experiment on active learning in mathematics. *Economic Letters* 156:172–175.

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

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

Brunetti, Aymo, Konstantin Büchel, Martina Jakob, Ben Jann, and Daniel Steffen. 2023. Inadequate teacher content knowledge and what could be done about it: Evidence from El Salvador. *Journal of Development Effectiveness* :1–24.

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

Callaway, Brantly and Pedro HC Sant'Anna. 2021. Difference-in-differences with multiple time periods. *Journal of Econometrics* 225 (2):200–230.

Camfield, Laura, Maren Duvendack, and Richard Palmer-Jones. 2014. Things you wanted to know about bias in evaluations but never dared to think. *IDS Bulletin* 45 (6):49–64.

Cornelius-White, Jeffrey. 2007. Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of educational research* 77 (1):113–143.

De Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2018. Double for nothing? Experimental evidence on an unconditional teacher salary increase in Indonesia. *The Quarterly Journal of Economics* 133 (2):993–1039.

Duflo, Esther, Rema Hanna, and Stephen P Ryan. 2012. Incentives work: Getting teachers to come to school. *American Economic Review* 102 (4):1241–78.

Garbarino, Sabine and Jeremy Holland. 2009. Quantitative and qualitative methods in impact evaluation and measuring results. Discussion Paper. University of Birmingham.

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

Harbour, Kristin E, Lauren L Evanovich, Chris A Sweigart, and Lindsay E Hughes. 2015. A brief review of effective teaching practices that maximize student engagement. *Preventing School Failure: Alternative Education for Children and Youth* 59 (1):5–13.

Jakob, Martina Saskia and Sebastian Heinrich. 2023. Measuring human capital with social media data and machine learning. University of Bern Social Sciences Working Papers 46, University of Bern.

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

Krumpal, Ivar. 2013. Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & quantity* 47 (4):2025–2047.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019a. Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. *The Quarterly Journal of Economics* 134 (3):1627–1673.

Mbiti, Isaac, Mauricio Romero, and Youdi Schipper. 2019b. Designing effective teacher performance pay programs: experimental evidence from tanzania. Tech. rep., National Bureau of Economic Research.

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

McKenzie, David. 2018. Can business owners form accurate counterfactuals? eliciting treatment and control beliefs about their outcomes in the alternative treatment status. *Journal of Business & Economic Statistics* 36 (4):714–722.

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

Miguel, Edward and Michael Kremer. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72 (1):159–217.

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. Teacher performance pay: Experimental evidence from india. *Journal of Political Economy* 119 (1):39–77.

NECTA. 2018. Format for standard four national assessment. Tech. rep., National Examinations Council of Tanzania.

———. 2020. Format for primary school leaving examinations. Tech. rep., National Examinations Council of Tanzania.

Patrinos, Harry A and Noam Angrist. 2018. Global dataset on education quality: A review and update (2000-2017). *World Bank Policy Research Working Paper* (8592).

Seidel, Tina and Richard J Shavelson. 2007. Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research* 77 (4):454–499.

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

Smith, Jeffrey, Alexander Whalley, and Nathaniel Wilcox. 2021. *Are participants good evaluators?* WE Upjohn Institute.

Snilstveit, Birte, Jennifer Stevenson, Daniel Phillips, Martina Vojtkova, Emma Gallagher, Tanja Schmidt, Hannah Jobse, Maisie Geelen, Maria Pastorello, and John Eyers. 2015. Interventions for improving learning outcomes and access to education in low- and middle- income countries: A systematic review. 3ie Systematic Review 24.

Sumra, Suleman, Sara Ruto, and Rakesh Rajani. 2015. Assessing literacy and numeracy in tanzania's primary schools: The uwezo approach. *Preparing the Next Generation in Tanzania* :47.

UN, United Nations. 2015. The 2030 agenda for sustainable development. New York: United Nations.

UNESCO. 2022. Pupil-teacher ratio in primary school in 2018, Tanzania. Published online: `https://data.worldbank.org`.

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

———. 2019. Teach. Our vision is to revolutionize how education systems track and improve teaching quality. World Bank Brief, published online: `www.https://www.worldbank.org/`.

Zizzo, Daniel John. 2010. Experimenter demand effects in economic experiments. *Experimental Economics* 13:75–98.

A Appendix: Additional results from experimental analysis

A.1 Baseline characteristics

Table A.1: Baseline characteristics

	Control	T1	T2	p-value
Panel 1: Teacher variables (N = 434)	(1)	(2)	(3)	(4)
Math score (percent correct)	77.390	78.523	77.606	0.644
	(0.874)	(0.914)	(0.994)	
Female	0.299	0.277	0.315	0.797
	(0.035)	(0.039)	(0.041)	
Age	38.203	38.654	36.984	0.285
	(0.676)	(0.816)	(0.757)	
Years since graduation	12.040	12.308	11.118	0.514
	(0.701)	(0.861)	(0.730)	
Panel 2: School variables (N = 219)				
Nr. of pupils that took SFNA	58.461	52.815	49.754	0.098
	(3.136)	(2.813)	(2.512)	
School PSLE avg. score (std)	-0.008	0.170	-0.096	0.323
	(0.103)	(0.119)	(0.146)	
Driving distance to district headquarters (h)	0.579	0.551	0.612	0.727
	(0.036)	(0.047)	(0.061)	
Nr. of pupils per class	43.574	39.755	40.377	0.309
	(2.022)	(1.540)	(2.037)	
Panel 3: Pupil variables (N = 10,101)				
Pupil math score (std)	-0.034	0.023	0.031	0.730
	(0.060)	(0.064)	(0.071)	
Pupil avg. score (std)	-0.007	0.028	-0.017	0.912
	(0.077)	(0.073)	(0.088)	
Pupil passed math exam	0.656	0.671	0.679	0.812
	(0.023)	(0.027)	(0.028)	
Pupil passed exam	0.764	0.788	0.742	0.544
	(0.026)	(0.027)	(0.033)	
Pupil scored A or B in math	0.390	0.422	0.421	0.603
	(0.025)	(0.027)	(0.029)	
Pupil scored A or B on avg.	0.359	0.378	0.371	0.901
	(0.031)	(0.031)	(0.035)	
Female pupil	0.523	0.512	0.504	0.293
	(0.008)	(0.009)	(0.009)	

Notes: Columns (1) - (3) report the mean for different covariates by experimental group (standard errors in parentheses). Column (4) reports the p-value of the F-test for differences in means across groups. Pupil baseline tests scores are taken from the *Standard Four National Examination* (SFNA), administered to all pupils in grade 4. School-level test scores from the *Primary School Leaving Examination* (PSLE), administered in grade 7, are used to assess the initial quality of the school.

A.2 Attrition at endline

Table A.2: Attrition of teachers at endline by experimental group

	All teachers		Targeted teachers		Peer teachers	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Pedagogy	0.006	0.004	0.032	0.039	-0.025	-0.036
	(0.038)	(0.037)	(0.056)	(0.055)	(0.054)	(0.054)
T2: Pedagogy & Content	0.057	0.046	0.044	0.041	0.064	0.042
	(0.046)	(0.047)	(0.059)	(0.061)	(0.065)	(0.065)
Baseline score	-0.014	0.004	-0.029	-0.017	-0.001	0.015
	(0.018)	(0.019)	(0.033)	(0.033)	(0.025)	(0.027)
Avg. attrition rate	0.151	0.151	0.146	0.146	0.156	0.156
Observations	434	434	219	219	215	215
Adjusted R^2	0.010	0.020	0.012	0.016	0.008	0.018
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Linear probability model estimating the impact of the treatments on attrition probability. Estimates reported for *all teachers* in columns (1) and (2), for *targeted teachers* in columns (3) and (4), and for *peer teachers* in columns (5) and (6). *Teacher level controls* include sex, age, and years since graduation. Huber-White robust standard errors in parentheses.
+ p<0.10, * p<0.05, ** p<0.01.

Table A.3: Attrition of pupils between SFNA 2018 and PSLE 2021 by experimental group

	Attrition	
	(1)	(2)
T1: Pedagogy	-0.011	-0.004
	(0.015)	(0.011)
T2: Pedagogy & Content	-0.011	-0.008
	(0.017)	(0.013)
Pupil baseline math score		-0.055**
		(0.007)
Observations	12657	11991
Adjusted R^2	0.018	0.044
Controls	No	Yes
Stratum fixed effects	Yes	Yes

Notes: Linear probability model estimating the impact of the treatments on attrition rates. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects) and number of pupils, and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01.

A.3 Robustness checks for main effects at the student level

Table A.4: Robustness checks for effects on students' math scores

	Standardized				Scored A or B			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.11^{+}	0.11^{+}	0.13*	0.14*	0.05*	0.05*	0.05*	0.06*
	(0.06)	(0.06)	(0.06)	(0.06)	(0.02)	(0.02)	(0.02)	(0.02)
Pupil baseline math score	0.47**	0.32**	0.33**	0.33**	0.12**	0.08**	0.08**	0.08**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	10101	10101	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.25	0.27	0.29	0.30	0.15	0.16	0.17	0.18
Pupil Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
School Controls	No	No	Yes	Yes	No	No	Yes	Yes
Teacher Controls	No	No	No	Yes	No	No	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized math scores in all models. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils, and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

Table A.5: Program effect on the average score of pupils across subjects

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.082	0.121*	0.053*	0.069**	0.000	0.009
	(0.059)	(0.054)	(0.025)	(0.023)	(0.019)	(0.018)
Pupil baseline avg. score	0.499**	0.306**	0.175**	0.105**	0.152**	0.095**
	(0.021)	(0.024)	(0.010)	(0.011)	(0.009)	(0.010)
Mean of dep. variable	-0.015	-0.015	0.230	0.230	0.798	0.798
Observations	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.272	0.325	0.200	0.250	0.169	0.193
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

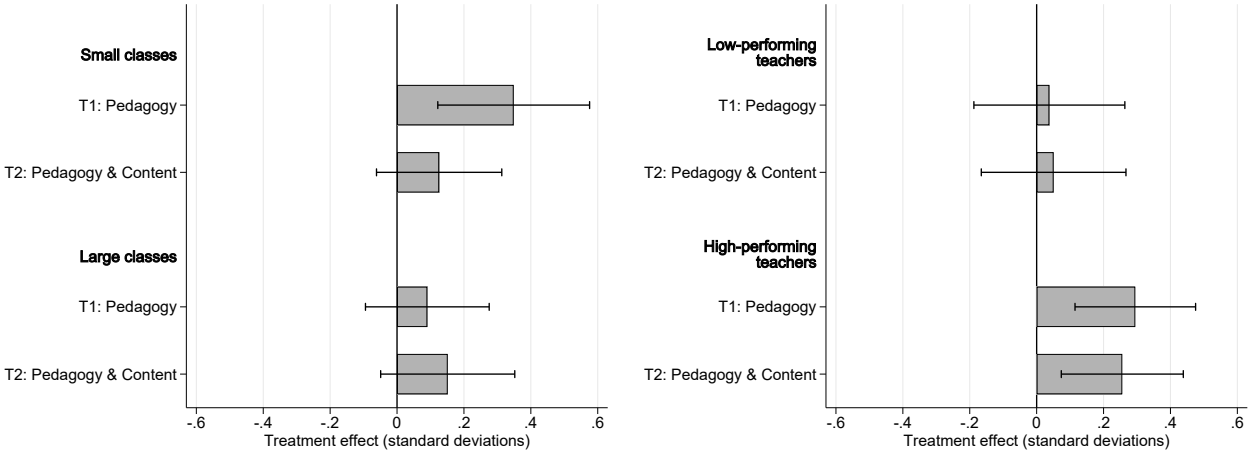
Notes: The *dependent variable* is pupils' standardized average score (across all subjects) for columns (1) and (2), a binary variable indicating whether a student's average score was A or B (3) and (4), and a binary variable indicating whether a pupil passed the exam for columns (5) and (6). *Pupil baseline math score* is a pupil's score in the SFNA exam administered in grade 4. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils in grade 4 and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

A.4 Effect heterogeneity and spillovers at the student level

Table A.6: Estimates for cascading effects on the math score of pupils

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Pedagogy	0.066	0.087	0.013	0.016	0.017	0.027
	(0.059)	(0.055)	(0.011)	(0.011)	(0.025)	(0.024)
T2: Pedagogy & Content	-0.011	-0.012	0.009	0.009	0.006	0.006
	(0.056)	(0.055)	(0.011)	(0.012)	(0.024)	(0.023)
School PSLE avg. score (std)	0.077*	0.080*	0.016**	0.017**	0.033**	0.033**
	(0.032)	(0.031)	(0.006)	(0.006)	(0.013)	(0.012)
School SFNA avg. score (std)	0.132**	0.125**	0.022**	0.022**	0.047**	0.045**
	(0.032)	(0.032)	(0.006)	(0.006)	(0.013)	(0.014)
T2 − T1	-0.077	-0.099	-0.004	-0.007	-0.012	-0.021
	(0.064)	(0.066)	(0.013)	(0.013)	(0.027)	(0.028)
Mean of dep. variable	-0.000	-0.000	0.075	0.075	0.368	0.368
Observations	15023	15023	15023	15023	15023	15023
Adjusted R^2	0.073	0.081	0.035	0.040	0.053	0.060
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized SFNA math score for columns (1) and (2), a binary variable indicating whether a pupil scored A or B (highest grades) in math for columns (3) and (4), and a binary variable indicating whether a pupil passed the math exam for columns (5) and (6). Controls include *pupil-level controls* for sex, *teacher-level controls* for sex, age, and math performance at baseline, and *school-level controls* for the number of pupils in grade 4 as well as each school's average SFNA and PSLE score in 2018. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.



(a) By class size (proxy)

(b) By teacher performance

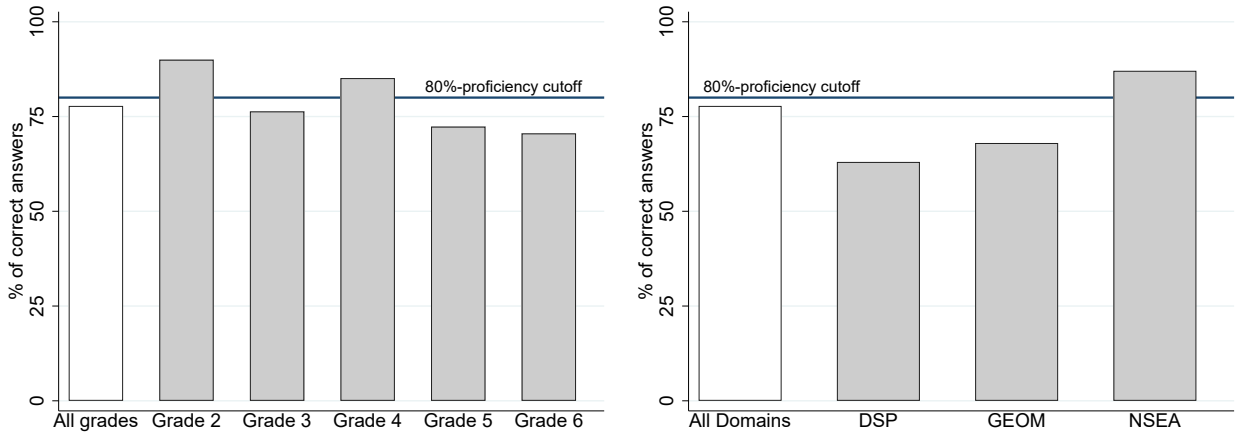
Figure A.1: Heterogeneity in treatment effects on students' math scores by class size and teacher mathematical content knowledge at baseline.
Groups are split at the median of class size and teacher performance. 90 percent confidence intervals shown.

Table A.7: Effect heterogeneity along attributes of pupils and teachers

Covariate:	Pupils' score		Female pupil		Teacher score		Class size	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.11^{+}	0.14*	0.11^{+}	0.15*	0.11^{+}	0.14*	0.11^{+}	0.14*
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Covariate	0.45**	0.33**	-0.03	-0.03	-0.06	-0.05	0.15	0.11
	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.06)	(0.13)	(0.15)
Treatment × Covariate	0.02	-0.00	-0.05	-0.07	0.14^{+}	0.12	-0.27	-0.17
	(0.04)	(0.04)	(0.06)	(0.06)	(0.07)	(0.07)	(0.17)	(0.16)
Observations	10101	10101	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.25	0.30	0.25	0.30	0.25	0.30	0.25	0.30
Teacher controls	No	Yes	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized math scores in all models. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils, and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

A.5 Descriptive statistics on teacher content knowledge



(a) Average scores by grade level of question

(b) Average scores by domain

Figure A.2: Math proficiency of teachers prior to the project

The assessment featured 50 items covering the math curriculum of Tanzanian primary schools (grades 2–6) and was administered in November 2019. Participants are either *targeted teachers* (N=219) or *peer teachers* (N=215) nominated for the evaluation study by public primary schools in Siha, Karatu, Mbulu DC, and Mbulu TC. Note that the sample is neither representative for Tanzanian teachers nor for teachers in the study regions.

A.6 Additional results for program effects at the teacher level

We use the following equation to estimate intermediate effects on teachers:

$$Y_{isk} = \beta_1 T1_s + \beta_2 T2_s + \beta_3 Peer_i + \beta_4 T1_s \times Peer_i + \beta_5 T2_s \times Peer_i + X_i^{'}\gamma + \phi_k + \epsilon_{isk}, \qquad \text{(A.1)}$$

where Y_{isk} is a teacher's math score after the intervention, $T1_s$ indicates if the teacher's school was assigned to the PEDAGOGY intervention, $T2_s$ represents if a teacher's school was in the PEDAGOGY & CONTENT group, $Peer_i$ indicates if the teacher was only a peer teacher rather than being directly targeted, and $T1_s \times Peer_i$ and $T2_s \times Peer_i$ are interaction terms capturing if treatment effects for peer teachers are different from those on directly targeted teachers. Finally, $X_i^{'}\gamma$ is a vector of teacher-level controls for sex, age and baseline score, ϕ_k are strata fixed effects, and ϵ_{isk} captures the error term.

Table A.8: Main estimation results for program effects on the math score of teachers

Dependent	Overall				NSEA		GEOM + DSP	
variable:	%		Standardized		Standardized		Standardized	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
T1: Pedagogy	0.40	0.47	0.03	0.04	0.02	0.04	0.12	0.10
	(1.37)	(1.40)	(0.10)	(0.11)	(0.12)	(0.12)	(0.11)	(0.11)
T2: Pedagogy & Content	1.95	2.00	0.15	0.15	0.22^{+}	0.22^{+}	0.04	0.05
	(1.31)	(1.33)	(0.10)	(0.10)	(0.11)	(0.11)	(0.11)	(0.11)
Peer teacher	-2.92*	-2.43^{+}	-0.22*	-0.19^{+}	-0.19^{+}	-0.15	-0.09	-0.07
	(1.33)	(1.34)	(0.10)	(0.10)	(0.12)	(0.12)	(0.10)	(0.10)
T1 × Peer teacher	2.56	2.39	0.19	0.18	0.26	0.23	-0.06	-0.05
	(2.10)	(2.07)	(0.16)	(0.16)	(0.19)	(0.18)	(0.16)	(0.16)
T2 × Peer teacher	-0.59	-0.35	-0.05	-0.03	0.01	0.02	-0.15	-0.13
	(1.99)	(2.00)	(0.15)	(0.15)	(0.18)	(0.18)	(0.17)	(0.17)
Baseline score	10.01**	9.54**	0.76**	0.73**	0.68**	0.64**	0.74**	0.72**
	(0.56)	(0.59)	(0.04)	(0.04)	(0.06)	(0.06)	(0.03)	(0.04)
T2 − T1	1.55	1.53	0.12	0.12	0.20	0.18	-0.07	-0.05
	(1.48)	(1.51)	(0.11)	(0.11)	(0.13)	(0.13)	(0.11)	(0.11)
Observations	368	368	368	368	368	368	368	368
Adjusted R^2	0.62	0.63	0.62	0.63	0.48	0.49	0.58	0.59
Teacher controls	No	Yes	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is the share of correct answers for columns (1) and (2), standardized test scores for columns (3) and (4), standardized test scores on NSEA (numbers sense and elementary arithmetic) items for columns (5) and (6), and standardized test scores on GEOM (geometry and measurement) and DSP (data, statistics and probability) items for columns (7) and (8). Main treatment effects are reported for *targeted teachers*, i.e. teachers directly exposed to the treatments. *Teacher level controls* include sex, age, and years since graduation at baseline. Huber-White robust standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01.

Table A.9: Heterogeneity in program effects on teachers' mathematics performance

Covariate:	Baseline score		Age		Female	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Pedagogy	0.031	0.032	0.053	0.046	0.015	0.020
	(0.109)	(0.112)	(0.104)	(0.106)	(0.119)	(0.119)
T2: Pedagogy & Content	0.128	0.131	0.136	0.147	0.104	0.099
	(0.092)	(0.094)	(0.098)	(0.101)	(0.114)	(0.114)
Covariate	0.730**	0.712**	0.002	0.001	-0.131	-0.169
	(0.059)	(0.060)	(0.008)	(0.017)	(0.145)	(0.149)
T1 × Covariate	-0.115	-0.112	-0.013	-0.011	0.049	0.121
	(0.115)	(0.119)	(0.011)	(0.012)	(0.266)	(0.286)
T2 × Covariate	-0.341**	-0.349**	-0.016	-0.020+	0.160	0.177
	(0.117)	(0.121)	(0.011)	(0.012)	(0.235)	(0.247)
Observations	368	368	368	368	368	368
Adjusted R^2	0.625	0.637	0.617	0.625	0.629	0.634
Teacher controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is teachers' standardized test scores in all models. *Heterogeneity* is estimated along teachers' baseline score in columns (1) and (2), teachers' age in columns (3) and (4), and teachers' sex in columns (5) and (6). Main effects are reported for *targeted teachers*, i.e. teachers directly exposed to the treatments. Age is centered to have a mean of 0 for targeted teachers, and baseline scores are standardized to have a mean of 0 and a standard deviation of 1 for targeted teachers. Estimates for *Peer teacher*, *Peer teacher* × *Treatment*, *Peer teacher* × *Covariate* and *Peer teacher* × *Treatment* × *Covariate* not shown. *Teacher level controls* include sex, age, and years since graduation at baseline. Huber-White robust standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01.

B Appendix: Difference-in-differences analysis

To observe potential spillover effects over a long time horizon, we conduct a multi-year ex-post analysis based on school-level data for both grade 7 and grade 4 students. The Primary School Leaving Examination (PSLE) for seventh graders has been conducted on a yearly basis since 2013, while the Standard Four National Assessment (SFNA) assessment for fourth graders was launched in 2015. Combining the publicly available national examination data with the NGO documentation on the program implementation allows us to trace how tests scores in program schools evolve relative to test scores in schools that did not participate in the teacher training program. As only one teacher (or a very small group of teachers) per school was invited to participate in the program, and selected teachers were then instructed to organize knowledge sharing activities with their colleagues, this comes close to an estimation of cascading effects. To be precise, it provides an upper bound for these effects, given that a small share of students should have been taught by directly targeted teachers.

With these considerations in mind, we estimate cascading effects associated with the program using

$$Y_{st}^{Std} = \beta_1 \, Treatment_{st} + \lambda_s + \phi_t + \epsilon_{st} \quad \text{for } Grade \in \{4, 7\}, \tag{B.1}$$

where Y_{st}^{Std} represents the average test score in math of school s in year t for either grade 4 (SFNA) or grade 7 (PSLE), *Treatment* indicates whether one or several teachers from a school participated in the training on the new teaching methods and is set to 1 for a given year t and later years if school s was part of the program in year t (and to 0 otherwise), λ_s are school level fixed effects, ϕ_t are year fixed effects, and ϵ_{st} is the error term.

This corresponds to a standard two-way fixed effects estimator (TWFE). To assess the robustness of the difference-in-differences analysis, the standard TWFE-estimates are compared to results obtained from an alternative difference-in-differences estimator proposed by Callaway and Sant'Anna (2021). As a control group, we use both never and not yet treated units. In all models, the comparison group consists of all schools from the three Tanzanian regions – Arusha, Kilimanjaro, and Manyara – where the project was implemented.

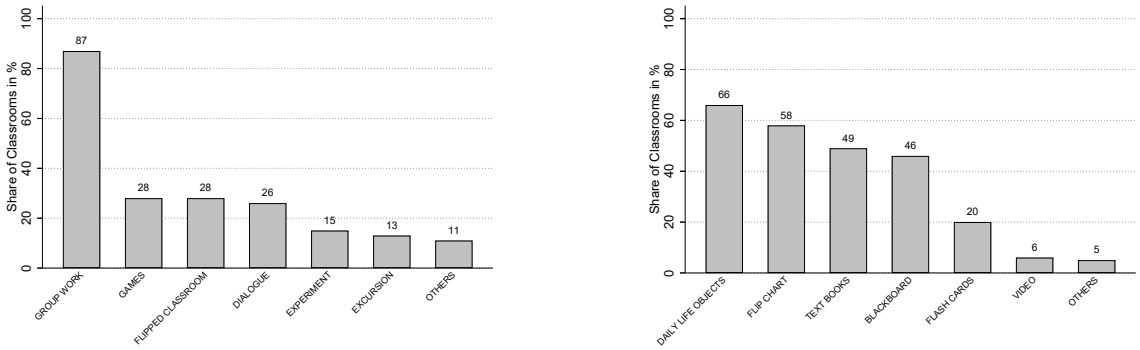
Results are presented in Table B.1. Across all models, effects are close to zero and insignificant.

Table B.1: School level difference-in-differences estimates for cascading effects, 2013–2019

	SFNA (grade 4)			PSLE (grade 7)		
	TWFE	CS		TWFE	CS	
		NE	NY		NE	NY
	(1)	(2)	(3)	(4)	(5)	(6)
ATT	0.032	-0.039	-0.033	-0.028	0.008	0.008
	(0.032)	(0.038)	(0.038)	(0.022)	(0.028)	(0.027)
Observations	11379	7168	7168	14954	11470	11470
Adjusted R^2	0.176			0.253		

Notes: The *dependent variable* are standardized test scores at the school level in all models. Effects in the standard TWFE model are compared with estimates obtained through the approach proposed by Callaway and Sant'Anna (2021), labeled as "CS". The presented CS coefficients stem from a comparison with *never treated* units (NE) or *not yet* treated units (NY). As the CS panel estimator does not take into account schools with incomplete data and always treated schools, they are based on a more restricted sample. To account for schools with incomplete data, results were also compared with a cross-sectional CS estimator and remain very similar (not shown). Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01.

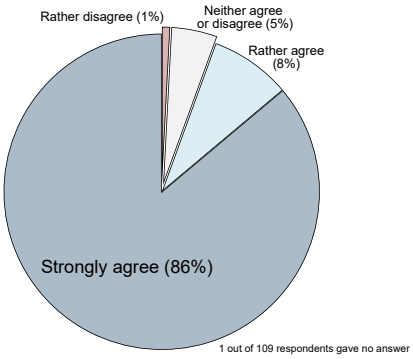
C Appendix: Classroom observations and opinion survey



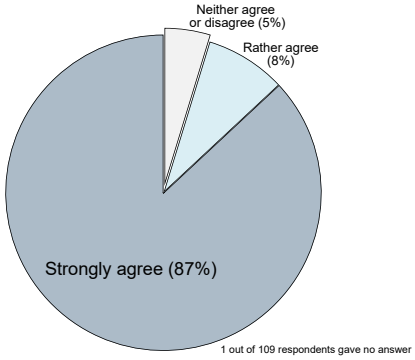
(a) Involvement of pupils

(b) Use of teaching aids

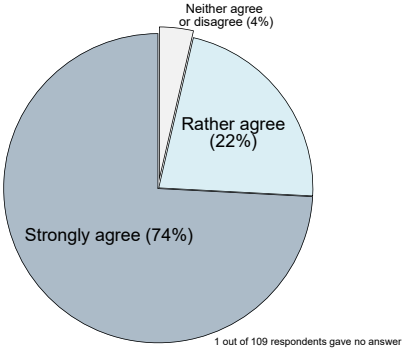
Figure C.1: Observed teaching techniques in treatment schools.
The data was collected by government employed Quality Assurance Officers in 112 out of 130 program schools.



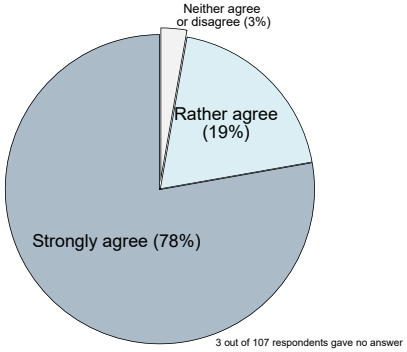
(a) Treated teachers: The project improved my math knowledge.



(b) Treated teachers: The project improved my teaching strategies.



(c) Treated teachers: The project improved the math skills of my pupils.



(d) Peer teachers: The project improved the math skills of my pupils.

Figure C.2: Perceived impact on teachers' *content knowledge in math*, their *teaching strategies* and their *students' math skills* as reported by the participants.
The treatment group includes 130 teachers, whereof 109 attended data collection, while the peer group includes 130 teachers, whereof 107 attended data collection.

D Appendix: Exemplary quotes from semi-structured interviews

Table D.1: Exemplary quotes from the semi-structured interviews conducted with SITT participants, SITT-D participants, peer teachers, and officials, part 1.

Group	General impression of SITT	Impact on math skills	Impact on teaching	Impact on pupils	Comparison with other educational programs
SITT	*"I really appreciate the SITT program, because it changed the way I deliver material to the classroom. [...] Thanks to SITT, I can use participatory methods that encourage pupils to contribute more actively."*	*"I understand mathematics very well. My main problem is how to teach it to the pupils. SITT showed me new ways in how to teach in the classroom. Concerning math skills, I gained some new ideas from the facilitators during the workshops."*	*"SITT helped me to involve kids in preparing teaching aids, and this helps the kids to remember the material better. [...] Another thing is that teachers are no longer working individually but together as a team. Pupils and teachers also came closer, you now find kids asking for the help of teachers."*	*"My knowledge increased and the way of teaching mathematics to my students improved so that my students learn better."*	Not discussed with SITT participants.
SITT-D	*"SITT is really good. It helped me so much. Before SITT, I was afraid to teach math. After participating in this program, I feel comfortable teaching math."*	*"There is a change in my math proficiency, because I use the computer with the 'Kolibri' learning software."*	*"SITT changed me quite a lot. Now I engage children more actively in my lessons. Instead of narrating like a radio, I teach practically."*	*"The program probably helps the students. When I use SITT methods they like it and they learn better."*	Not discussed with SITT-D participants.
Peers	*"SITT is useful to us, because it helps our pupils to prepare teaching aids [...] and it makes teaching more learner-centered. SITT will change our school, everybody loves it."*	Not discussed with peer teachers.	*"The SITT program has improved my teaching much, because it remembered me to use teaching aids and participatory methods."*	*"Pupils enjoy when we teach them according to SITT. That makes them understand more easily."*	Not discussed with peer teachers.
Officials	*"SITT is nice and very good for the teachers. Not only for the teaching aids and teaching materials but also for the technology. The teachers are learning through the computer and software."* *"I agree with my colleague. On WhatsApp, I observe what the teachers are sharing. It is really impressive and the teachers are enjoying it."*	Not discussed with officials.	*"During my school visits, I observed that SITT teachers have a different teaching approach. For instance, they try to use teaching aids and participatory methods."*	*"For now, it is difficult to say how large the effect of SITT is, because the pupils have been taught by several teachers between standard 1 and standard 7. So, I am not sure by how much SITT helps the performance of kids."*	*"I remember a program phasing out in 2012 that offered an in-service training. It was introduced and supported by UNICEF. [...] It was considered too burdensome by the teachers so they didn't work on it properly. [...] The program ended and the results were disappointing. For the case of SITT, the peer-sharing within school works better. Also the idea of model lessons helps. And SITT's unique participatory approach motivates pupils and makes them like mathematics more."*

Sources of quoted statements: Interviewees in Mbulu DC (×4), interviewees in Mbulu TC (×4), interviewees in Karatu (×3), interviewees in Siha (×4).
SITT refers to the group receving only the PEDAGOGY intervention, and SITT-D to the group that additionally recevied the laptops for content revisions, i.e. PEDAGOGY + CONTENT.

Table D.2: Exemplary quotes from the semi-structured interviews conducted with SITT participants, SITT-D participants, peer teachers, and officials, part 2.

Group	Feedback: Workshops	Feedback: Laptop/Kolibri	Feedback: Cascading	Relevance of evaluation	Additional remarks
SITT	*"I liked the training as it made me a better teacher. I also appreciated the change in environment from Mbulu to Arusha and the good service."*	Not discussed with SITT participants.	*"The perspective of my colleagues was a problem. I called a meeting, and they agreed to my proposal. But once I asked them to join team teaching, most of them said 'Now, I have no time'. At other schools it is similar."*	Not discussed with SITT participants.	**About Covid-19 and the future:** *"We temporarily closed schools due to Covid in 2020. Still, we used SITT to improve our teaching and that is why we achieve a good performance in our school. I ensure that we will keep it and improve even more."*
SITT-D	*"I liked the workshop very much, but I was disappointed that the additional meetings for SITT-D in 2019 were canceled [because of Covid-19]."*	*"The laptop and learning software are very useful. Kolibri helps mathematics teachers to be up to date. We use it to refresh our knowledge before teaching a certain topic. It makes us comfortable."*	*"We created a timetable to plan the model lessons and team teaching. Now, I see my colleagues using teaching aids. They like it and cooperate."*	Not discussed with SITT-D participants.	**Training intensity:** *"It would be good to have more than a 5-day workshop to have additional time to learn and share with teachers from other districts."*
Peers	Not discussed with peer teachers.	Not discussed with peer teachers.	*"Once our colleague shared their SITT-knowledge, we agreed together to have team teaching. [...] Around ninety percent appreciate it. [...] We will continue to use the techniques that the SITT project introduced."*	Not discussed with peer teachers.	**The cascading approach:** *"We assessed each other on how we conduct model lessons and discussed it during meetings. But there are some challenges: Not all teachers were eager to participate in the knowledge sharing activities."*
Officials	Not discussed with officials.	Not discussed with officials.	Not discussed with officials.	*"It is important to conduct an evaluation so that the implementers get feedback on what they are doing and to see whether it is useful or not. Spending money on an evaluation is necessary."*	**The relevance of evaluations:** *"It is very important to do the evaluation and to understand whether the program delivers or not."*

Sources of quoted statements: Interviewee in Mbulu DC (×1), interviewee in Mbulu TC (×2), interviewee in Karatu (×4), interviewee in Siha (×4).
SITT refers to the group receving only the PEDAGOGY intervention, and SITT-D to the group that additionally recevied the laptops for content revisions, i.e. PEDAGOGY + CONTENT.