Journal of Development Economics 181 (2026) 103742

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Regular article

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

Martina Jakob [a,*], Konstantin Büchel [b,c], Daniel Steffen [d], Aymo Brunetti [b]

[a] University of Zurich, Switzerland
[b] University of Bern, Switzerland
[c] Youth Impact, Botswana
[d] Lucerne University of Applied Sciences and Arts, Switzerland

ARTICLE INFO

Dataset link: https://doi.org/10.7910/DVN/DH3PZA

JEL classification:
I21
J24
O15
C93

Keywords:
Education quality
Teacher training
Participatory teaching
Teacher content knowledge
Computer-assisted learning
Development economics

ABSTRACT

While participatory teaching methods have been shown to be more successful than traditional rote learning in high-income countries, it is 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 0.13σ. The additional provision of laptops with a learning software allowing a random subset of teachers to refresh their content knowledge did not yield further learning gains for students. We also find limited evidence of spillover effects on indirectly exposed teachers and their students, even though knowledge-sharing activities were a key component of the program. Complementary findings from participant surveys and interviews suggest that the program was highly appreciated by different stakeholders, but that participants were unable to assess its impact along different dimensions, giving equally positive evaluations of its successful and its unsuccessful elements.

1. Introduction

Only 4 percent of students in low-income countries, compared to 95 percent in high-income countries, attain minimum literacy skills towards the end of primary school (World Bank, 2018). To narrow the global learning gap, it is critical to reconsider the strategies that teachers in developing countries use in the classroom. While schools in high-income countries have increasingly adopted participatory pedagogical approaches with high levels 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 sees student engagement as a vital component of effective teaching, a view that is corroborated by extensive evidence from high-income countries (e.g., Cornelius-White, 2007; Seidel and Shavelson, 2007; Harbour et al., 2015). However, it is not clear whether 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 of inadequate subject matter mastery among many teachers in low- and middle-income countries (e.g., Bold et al., 2017a; Brunetti et al., 2023), it remains an open question whether teacher content knowledge represents a binding constraint for the effectiveness of pedagogical interventions.

To address these questions, we conducted a randomized controlled trial (RCT) with 440 math teachers and more than 25,000 students

☆ Martina Jakob and Konstantin Büchel share first authorship. The authors acknowledge generous funding by the IMG Stiftung. This study received IRB approval from the Faculty of Business, Economics and Social Sciences of the University of Bern on November 4, 2019 (serial number: 122019). This RCT was registered as AEARCTR-0004959. The registry is available at: https://www.socialscienceregistry.org/trials/4959. We would like to thank Donatian Marusu, Judith Sarapion, Ana Kallonga, Stephen Masunga, Diana Zacharia, Staphord Chalamaganza, and Peter Samwel for the superb coordination of the field work. We are also very grateful to Mauricio Romero, Ben Jann, and several anonymous reviewers for their valuable comments on this manuscript.

* Corresponding author.
E-mail addresses: martina.jakob@econ.uzh.ch (M. Jakob), konstantin.buechel@outlook.com (K. Büchel), dani.steffen@hslu.ch (D. Steffen), aymo.brunetti@unibe.ch (A. Brunetti).

https://doi.org/10.1016/j.jdeveco.2026.103742
Received 7 January 2025; Received in revised form 22 January 2026; Accepted 25 January 2026
Available online 5 February 2026

from 220 schools in Tanzania. With an average of 51 students per teacher and a persistent shortage of classrooms and teaching materials, Tanzania faces resource constraints that are typical for many education systems in low-income countries (UNESCO, United Nations Educational, Scientific and Cultural Organization, 2020). 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 everyday life, craft teaching aids from readily available local materials, and collaborate in teams to handle large classrooms and share teaching strategies. After the initial five-day workshop, all teachers were invited to three biannual refresher meetings to review concepts and discuss implementation issues. The intervention targeted grade 6 to 7 teachers and leveraged a cascading model where participants were instructed to share their knowledge with their colleagues through dedicated activities. 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 attended additional sessions to familiarize themselves with the program and discuss their progress. Both treatment versions were delivered by Helvetas, one of Switzerland's largest development organizations, which has been operating in Tanzania for over 50 years.

To estimate the impact of the program on student learning, we scraped student-level data from standardized assessments published by the National Examinations Council of Tanzania (NECTA). These nationwide assessments are administered annually to all students in grade 4 (Standard Four National Assessment, SFNA) and grade 7 (Primary School Leaving Examinations, PSLE). Our study design enables us to analyze the direct effect on students taught by treated 6th- and 7th-grade teachers using PSLE data, as well as spillover effects on students taught by indirectly exposed peer teachers using SFNA data. Since participating teachers also instruct subjects other than math, we can further estimate spillovers to these other subjects. We also use data from our own math assessment with teachers to study intermediate effects on participating teachers and their peers. To better understand the mechanisms behind potential effects, we complemented the experimental data with classroom observations, surveys, and in-depth interviews.

We report four sets of findings. First, the training in participatory pedagogy successfully improved students' test scores by 0.13σ ($p = 0.045$) 1.5 years later, and the share of students with top grades increased by 5 percentage points from 12 to 17 percent ($p = 0.035$). The point estimate for pass rates is also positive, but not statistically significant ($p = 0.19$). This is also consistent with our complementary non-experimental data showing that teachers applied a wide range of the participatory pedagogical strategies and expressed great enthusiasm for the program. The reported effects place the program in the top 30% of impact estimates for math interventions reviewed in Evans and Yuan (2022), and are particularly remarkable given that we used data from official national tests that were not specifically tailored to the intervention.[1] To further contextualize these results, we compare them with two other recent large-scale experiments in Tanzania analyzing the same government assessments: Mbiti et al. (2019), who study school inputs and teacher incentives, and Mbiti et al. (2023), who test teacher performance pay schemes. Both studies find substantial effects on researcher-administered tests but null results on the nationally standardized PSLE assessments. Our pedagogical intervention, by contrast, demonstrates measurable impacts on this nationally standardized examination.

Second, students who were taught by teachers equipped with laptops and CAL software to refresh their content knowledge did not outperform students whose teachers only participated in the pedagogical training program. The point estimate for the difference between teacher training with and without CAL software is small and not statistically significant. While the teachers who received a laptop with CAL software improved their understanding of concepts related to the subdomain of number sense and arithmetic by 0.18σ ($p = 0.077$), the effect on an overall math proficiency score is statistically insignificant ($p = 0.135$). The average teacher scored 78 percent correct answers at baseline, suggesting that many teachers were already sufficiently proficient in their subject before the intervention.[2] A heterogeneity analysis shows that the CAL-based refresher was significantly more effective for teachers with low content knowledge at baseline, strengthening the hypothesis that the good proficiency at baseline mitigated the impact of providing access to the CAL software.

Third, we find limited evidence for spillovers on indirectly exposed teachers and students in treatment schools, even though the program was specifically designed to produce such externalities. While trained teachers and their peers reported engaging in cascading activities such as model lessons and peer learning groups, and peer teachers improved their math skills by 0.18σ ($p = 0.05$), the estimated effect on their students is small and insignificant (0.05σ, $p = 0.32$). However, we do find some indication of spillovers to other subjects, with effects on the average student score across non-math subjects of up to 0.09σ ($p = 0.11$) and an increase in students scoring A or B of 5 percentage points ($p = 0.04$). This can be taken to suggest that teachers were able to transfer the pedagogical strategies they were instructed to use in their math classes to other subjects.

Fourth, we compare participants' beliefs about program impacts with the actual causal estimates from the RCT. This comparison suggests that participants' survey and interview responses are not 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 compelling evidence of such spillovers in our experimental data.

Our study ties into the literature on how teachers contribute to and may mitigate the global learning crisis (Kremer et al., 2013; Glewwe and Muralidharan, 2016; World Bank, 2018). While the role of teacher incentives and pay has been extensively studied (e.g., Muralidharan and Sundararaman, 2011; Duflo et al., 2012; De Ree et al., 2018; Mbiti et al., 2023), teacher performance depends not only on economic incentives but also on the range of instructional strategies available to teachers (e.g., Evans and Mendez Acosta, 2021; de Barros et al., 2021; Nourani et al., 2023). Accordingly, many countries invest heavily in teacher professional development. In the United States, teachers spend 19 days in training every year at an estimated cost of 18,000 USD per person (Jacob and McGovern, 2015). Similarly, in a large-scale survey across more than 40 countries, 95 percent of teachers reported participating in professional development activities in the past year (OECD, Organisation for Economic Co-operation and Development, 2019). Yet, despite the vast practical importance of such programs, research on

[1] Using external assessments addresses two major validity concerns inherent in typical educational experiments. First, it prevents researchers from tailoring assessments too closely to the intervention, thus preserving external validity. Second, it avoids differential test-taking effort in response to the treatment, thereby safeguarding internal validity.

[2] International data from almost identical assessments provide further context: 224 primary school teachers in El Salvador scored on average 47 percent correct answers, while the median Swiss teacher in a convenience sample of 16 participants answered 90 percent of the questions correctly (Brunetti et al., 2020). This comparison indicates that math competency among Tanzanian teachers in our sample is closer to that of Swiss rather than Salvadorian teachers.

their effectiveness has not kept pace.[3] In a recent review of teacher professional development (PD) interventions, Popova et al. (2022) conclude that "few PD programs are evaluated" and among those that are, "there is much more variation in effectiveness across teacher training programs than across education programs more broadly". The World Development Report 2018 echoes this view, noting that "a lot of teacher professional development goes unevaluated" and "most teacher training is ineffective, but some approaches work" (World Bank, 2018, p. 131). Particularly successful examples include training programs on scripted lessons (Piper et al., 2018; Gray-Lobe et al., 2022), teaching at the right level methodology (Banerjee et al., 2017), and targeted feedback from coaches (Cilliers et al., 2020, 2022a; Marinelli et al., 2023). Our study adds to this literature by showing that a five-day training in participatory pedagogy can induce teachers to restructure their classes and achieve higher learning gains for their students—even when classes are large and few teaching aids are readily available.

Beyond the direct effects on student learning in mathematics, we also report evidence of gains extending to other subjects taught by the same teachers. These cross-subject patterns contrast with findings for related pedagogical interventions: Structured pedagogy programs, which predominantly focus on subject-specific materials and content, tend to be highly effective in the targeted domain Evans and Popova (2016), but often fail to improve learning in other domains (e.g., Cilliers et al., 2022b; Buhl-Wiggers et al., 2023). By comparison, trainings that target broadly applicable pedagogical methods, such as the one evaluated in this study, usually generate smaller effects (Popova et al., 2022; Ganimian and Murnane, 2016), but appear to be more likely to produce improvements across multiple subjects (e.g., Cilliers and Habyarimana, 2023; Wolf et al., 2019). This suggests that the generalizability of teacher training depends on whether interventions emphasize transferable pedagogical strategies or subject-specific content, and implies that the full cost-effectiveness of general-pedagogy interventions may be underestimated. Our observed direct effects and cross-subject spillovers are achieved within a large-scale program targeting regular government teachers in public schools, conducted in collaboration with government officials—a setup where previous training programs often show null results (Loyalka et al., 2019; GEEAP, Global Education Evidence Advisory Panel, 2020). Promoting more engaging and transferable teaching strategies in low- and middle-income countries may thus be an essential element in the global quest for "inclusive and equitable quality education" (UN, United Nations, 2015).

Our paper also adds to a growing strand of literature studying the role of teacher content knowledge in the educational production function (e.g., Metzler and Woessmann, 2012; Bietenbeck et al., 2018; Brunetti et al., 2023). Our findings suggest that the majority of teachers in our sample demonstrate considerable subject mastery, and that providing a laptop with CAL software alongside the pedagogy training did not add measurable value for student learning. Even for the least proficient teachers, effects on content knowledge were too small to translate into meaningful improvements in student learning. Available estimates for content knowledge gains for CAL-based teacher training range from 0.13σ ($p = 0.16$) in Tanzania to 0.29σ ($p < 0.01$) in El Salvador (Brunetti et al., 2023), broadly aligning with reported effect sizes for CAL programs targeting students (Escueta et al., 2020). Given that a 1σ increase in teacher content knowledge yields only 0.09σ in annual student learning (Bau and Das, 2020; Metzler and Woessmann, 2012), substantially larger effects at the teacher level would be needed to meaningfully impact student outcomes.

We also contribute to the literature on treatment externalities. The canonical example of treatment externalities in education was documented by Miguel and Kremer (2004), where treating students with deworming pills produced large spillover effects on non-targeted children in nearby schools. Such externalities can drastically boost cost-effectiveness, motivating *cascading models*, where trained teachers transmit knowledge to untrained colleagues. Despite some concerns about quality dilution (e.g., Kerwin and Thornton, 2021; Romero et al., 2022), teacher professional development programs often employ cascade models for scaling (Orr et al., 2013; Popova et al., 2022).[4] A related technique prominently discussed in educational science involves teacher communities of learning (or practice): small groups who meet regularly to observe instruction, analyze teaching, and refine practice. Non-experimental research documents favorable results of such professional learning communities on teacher mindsets, teacher practice, and student achievement (e.g., Vescio et al., 2008; Christensen and Jerrim, 2025), but two recent experimental evaluations in Australia yield mixed evidence regarding student achievement (see Gore et al., 2021; Vaughan et al., 2023). Our results suggest that achieving measurable externalities via peer-to-peer cascading is challenging, likely because teachers need considerable high-quality exposure to new teaching strategies to effectively restructure their classes. Our cost-effectiveness calculations show that even small spillovers—below the detection threshold of typical education RCTs—would dramatically increase the effectiveness of the program, underscoring the importance of further research on peer-to-peer transmission mechanisms.

Finally, this paper contributes to the policy debate on the best methods to evaluate programs (Banerjee and Duflo, 2009). Despite the emphasis on causal inference methods within academia, rigorous impact evaluations are not always feasible, and methods like interviews and feedback surveys with project beneficiaries are often the norm in practice.[5] While these approaches provide valuable insights and complement experimental data, our findings underscore that they may be ill-equipped to assess program impact and distinguish between successful and less successful elements.

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 of developing countries. The massive expansion of schooling since the late 1990s has put considerable strain on schools throughout the country, resulting in shortages of teachers, classrooms, and teaching materials. As a result, the pupil–teacher ratio in primary schools exceeds 50 students per teacher (UNESCO, United Nations Educational,

[3] To assess how much attention pre- and in-service training of teachers has received in experimental research, we analyzed all AEA RCT registries published between 2013 and October 2024. Of the 1870 education-related trials registered in the AEA RCT registry, only 409 (22%) include the term "teacher", and 62 (3%) mention terms related to "teacher training" or "professional development" in the abstract (see Figure A.1 in the Online Appendix).

[4] Cascade models have been the traditional approach to teacher development in various countries, including Kenya (see Bett, 2016), South Africa (see Dichaba and Mokhele, 2012), or India (see Barrett, 2010), and international organizations provide toolkits on how to implement them (e.g., British Council, 2018). In a recent review on teacher professional development, Popova et al. (2022) identify cascading elements in 50 percent of the reviewed teacher training programs, and their binary indicator for a cascade-based designs was uncorrelated with metrics of program effectiveness.

[5] For example, USAID conducted impact evaluations for 11 percent of its projects between 2016 and 2022, with the highest share of impact evaluations (23%) in the education and social services sector (USAID, United States Agency for International Development, 2024). The share of impact evaluations is markedly lower at the Swiss Agency for Development and Cooperation, whose Evaluation and Corporate Controlling Unit classifies 5 percent of its evaluation reports between 2014 and 2024 as impact evaluations (SDC, Swiss Agency for Development and Cooperation, 2024). In a recent review of evaluation practices in Swiss development cooperation, the Parliamentary Control of the Administration (PCA, Parliamentary Control of the Administration, 2023) criticized the quality and use of evaluation reports as being inadequate for accountability purposes, while acknowledging that the evaluation standards in Switzerland are comparable to that in other countries.

Scientific and Cultural Organization, 2020), with one recent estimate placing the national average as high as 73 (UNICEF, 2024). In this context, the country has struggled to translate enrollment into learning. For example, about 60 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. Yet, 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 is implemented by Helvetas, one of the largest Swiss development organizations operating in Africa, Asia, Latin America, and Eastern Europe. Helvetas has been active in Tanzania for over 50 years, with projects in a wide range of areas, including agriculture, youth employment, and education. Following several years of piloting teacher professional development approaches, Helvetas, in collaboration with the Tanzanian Teachers' Union (TTU) and the Ministry of Education, launched the *Inclusive School-Based In-Service Teacher Training (SITT)* program in 2016. This large-scale program aims to transform pedagogy in Tanzanian classrooms. Prior to the experimental evaluation we discuss in this paper, the program (and its preceding pilots) had already been rolled out in 1400 schools throughout northeastern Tanzania.

The main objective of the SITT program is to promote a more *student-centered approach* to teaching. The training program focuses on four simple strategies to foster active student engagement. First, teachers learn how to prepare, facilitate, and monitor activities such as group work and peer teaching, where students take turns explaining concepts to the class. A specific focus is on ensuring the engagement of all and not just high-performing students in these activities. Second, teachers are taught how to incorporate games and practical examples from everyday life to make lessons more accessible and entertaining. Third, the training emphasizes the use of readily available local materials like wooden sticks, stones, and berries to address the scarcity of teaching resources. To strengthen the active role of students in the learning process, teachers are encouraged to engage students in crafting teaching aids. Finally, the program's rallying call "maksudi, maksudi", meaning "intentional, intentional" in Swahili, reminds teachers that every action—or lack thereof—is their deliberate choice, whether it is being well-prepared or engaging students effectively.

To facilitate *cascading*, teachers are encouraged to share their knowledge with all other teachers in their school through various collaborative activities. In particular, they are expected to invite their colleagues to model lessons showcasing the new teaching methods in action. Trained teachers are also asked to organize peer learning groups where colleagues reflect together on their experiences implementing these new pedagogical techniques in their classrooms.

During the experimental evaluation period, the intervention was supplemented with additional activities to address potential shortfalls in teachers' content knowledge. Half the teachers participated in an extended version of the SITT program, where they received a laptop equipped with a *computer-assisted learning* software. Learning materials used for this treatment version include video content and short quizzes in Swahili produced by Khan Academy and are delivered through Kolibri, an offline-first learning platform developed by Learning Equality. The instructional videos are 5 to 10 min long and grouped along three broad themes: *(i)* Number Sense and Elementary Arithmetic (NSEA, 80 videos), *(ii)* Geometry and Measurement (GEOM, 80 videos), and *(iii)* Data, Statistics and Probability (DSP, 11 videos). The videos are shared through a user-friendly interface and accompanied by short quizzes. Each quiz draws on a basis of roughly 20 items that are presented in random order. Upon submitting an answer, users receive instant feedback. The software tracks performance and awards badges of success for quizzes with at least five correct answers. Previous studies have shown that computer-assisted learning with Khan Academy is effective in improving test scores for both students (Büchel et al., 2022) and teachers (Brunetti et al., 2023).

The *delivery* of the SITT training consists of a five-day course institutionally embedded in government structures, with senior officials from the president's office, regional quality assurance, district education offices, and the Tanzania Teachers' Union presiding over openings and closings. The training program relies on a hands-on approach, where teachers work in groups to practice the program's student-centered pedagogy. During the training, each group prepares and delivers a math lesson, with other participants acting as students and providing feedback. Teachers further receive a comprehensive manual explaining the new teaching strategies with example activities. On the final day, they submit individual 'Action Plans' outlining how they will implement the new pedagogical strategies over the next six months. After the main workshop, teachers attend biannual two-day refresher meetings to reflect on how they applied their individual 'Action Plans' and they join a WhatsApp group to continuously share classroom experiences.

Table A.1 in the Online Appendix compares the program characteristics along the classification of top-performing teacher professional development interventions reviewed in Popova et al. (2022). In terms of content and delivery, the SITT program shares several characteristics of top performers, such as a multi-day face-to-face training event, a specific subject focus, lesson enactment during the training, a significant proportion of time allocated to practicing, and follow-up visits. However, it differs in organizational features, as participation carries no immediate implications for salary or promotion.

The intervention took place in 2020 and 2021, starting with an initial five-day training event in February 2020 for all 130 treatment teachers, followed by three refresher meetings over the next 18 months. Although this period coincided with global workplace and school closures due to the COVID-19 pandemic, Tanzania experienced one of the least restrictive lockdowns worldwide. While the average country mandated business closures or remote work for some or all sectors on 350 days, Tanzania did not introduce any such measures, along with only three other countries. Similarly, schools were fully closed for 223 days on average worldwide, but only for 75 days in Tanzania (Hale et al., 2021). Accordingly, student exposure to the new teaching techniques was substantial despite the disruptions caused by the pandemic.

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 of Karatu (Arusha), Siha (Kilimanjaro), and Mbulu DC and Mbulu TC (Manyara), where the program had not yet been implemented. Program regions broadly align with national averages across various indicators, including consumption, poverty, and household size, with a tendency towards somewhat better outcomes (Table B.1).[6] The most pronounced difference is the lower pupil–teacher ratio in program regions (i.e., 42:1) relative to the national average (i.e., 54:1). The implementing organization adopted a selection protocol similar to earlier implementation phases, excluding the best performing and the geographically least accessible schools in each district. Compared to other schools in Tanzania, the selected schools have a slightly higher math performance in grade 4. In terms of our main outcome, math performance in grade 7, they are statistically indistinguishable from the national average (Table B.2).

The experimental design allows us to identify *direct effects* on participating teachers and their pupils as well as *cascading effects* on peer teachers and their pupils. Specifically, selected schools nominated

[6] Note that there is considerable heterogeneity in program regions: Kilimanjaro performs above the national average on most socio-economic indicators, while Manyara often falls below, and Arusha is roughly aligned with national averages.

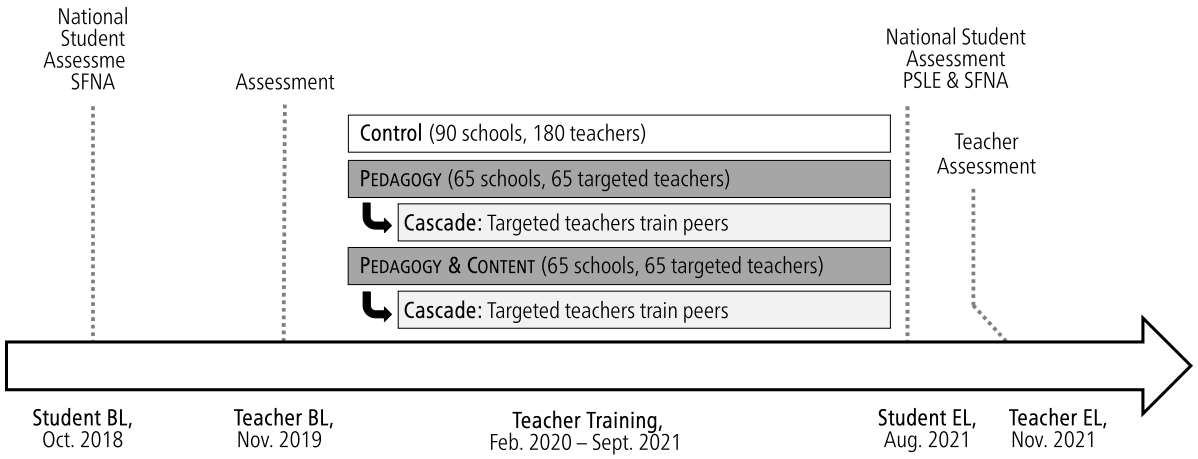


Fig. 1. Timeline of the study *Notes:* The main intervention event is a five-day workshop for all 130 treated teachers and was conducted in February 2020. Afterwards, teachers implemented the new strategies and shared them with their colleagues, participated in biannual meetings, and were visited by 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.

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 that *(i)* both teachers should instruct math, and that *(ii)* the targeted teacher should teach math to sixth-grade students in 2020 and to seventh-grade pupils in 2021. Targeted teachers usually had previous experience teaching grades 6 and 7, with 83 percent indicating that they typically teach math in these grades, and they signed an agreement with local education authorities committing to the teaching schedule required for evaluation purposes. 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 Fig. 1):

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

We refer to all 130 schools assigned to either the Pedagogy version (65 schools) or Pedagogy & Content version (65 schools) of the teacher training as the *treatment group*. 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., performance on the math assessment conducted in November 2019).

3.2. Data

We rely on nationally standardized tests to measure effects on students, and we conducted our own assessments to study intermediate effects on teachers. This experimental data is complemented by 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), taken in grade 7, and the *Standard Four National Assessment* (SFNA), taken in grade 4. These yearly assessments are administered to the entire student population in the respective grades and carry high stakes: failing SFNA can require pupils to repeat a grade, and passing PSLE is mandatory for admission to public secondary schools. 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 contains 25 math questions students need to answer in 90 minutes (NECTA, 2018, 2020); student-level results for both PSLE and SFNA examinations are publicly published as letter grades (A through E, where A is the top grade and E is the lowest) on the website of the National Examinations Council of Tanzania.

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. Students' PSLE scores can be merged with their SFNA scores from three years earlier (i.e., 2018) to establish a student-level baseline score. To assess spillover effects through *cascading*, the SFNA math scores of fourth-grade students in 2021 can be used, since these pupils were taught by peer teachers in the same school who were exposed to cascading activities. While these students were not necessarily taught by the peer teacher selected for study participation, this is inconsequential for studying pupils' learning outcomes, as cascading activities targeted all teachers in the school, and peer teachers did not have a special role in the intervention.

Since both PSLE and SFNA results are published online, we use web scraping to obtain the student-level data. Our final sample consists of 10,132 seventh graders to assess the direct effects of the programs and 15,073 fourth graders to estimate spillovers.

Our reliance on nationally standardized assessments designed and conducted by a governmental agency offers a key advantage over the typical educational field experiment, where no such independent data is available. A close alignment between the intervention and the assessments can compromise the validity of the results for two key reasons. First, researchers may–consciously or unconsciously–tailor the content of the assessments closely to the intervention, thereby limiting the external validity of the concepts they attempt to measure. Second, if participants perceive assessments as connected to the intervention, this may elicit differential test-taking effort between treatment and control groups and compromise the internal validity of the experiment. A prominent example comes from the pay-for-performance literature,

where assessments tend to be high-stakes situations for the treatment group, but low-stakes situations for the control group (for a discussion, see Mbiti et al., 2019, 2023). Similarly, socially motivated behavioral responses are plausible, such as increased test-taking effort in the treatment group due to gratitude for the intervention, or decreased effort in the control group due to disappointment at not being selected. Using data from national assessments addressed both concerns. A potential drawback is that high-stakes assessments can induce strategic behavior, such as the exclusion of low-performing students (Cilliers et al., 2021; Jacob, 2005), cheating (Singh, 2024), or teaching to the test (Jacob, 2005). However, as the stakes involved in Tanzania's national assessments are not related to the treatment status, potential strategic responses are unlikely to bias our experimental estimates. We can also exclude biases from grading, as the NECTA-appointed examiners who mark the assessments have no knowledge of students' treatment status.

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 2nd and 7th grade covering the domains of Number Sense & Elementary Arithmetic (NSEA, about 60%), Geometry & Measurement (GEOM, about 35%), and Data, Statistics, & Probability (DSP, about 5%). Assessments were administered as paper-and-pencil tests at regional meetings and had to be completed in 90 min.

Complementary non-experimental data. We collected three different types of quantitative and qualitative data to gain deeper insight into how switching to participatory pedagogy was viewed and put in practice by treated teachers. First, all teachers completed a short survey about their evaluations of the program and their perceptions of how it had impacted them and their students. The survey consisted primarily of single-choice questions that asked respondents to rate certain elements or indicate whether they agreed or disagreed with a given statement, but also included space for written feedback and suggestions. Surveys were administered during the endline math assessment to all teachers and tailored to the different experimental groups.[7] Second, to better understand how teachers incorporated the new methods into their teaching, quality assurance officers from the Ministry of Education conducted classroom observations of program participants' lessons. Starting with the Teach tool proposed by the World Bank (2019), we designed a classroom observation instrument and briefed government officials on how to conduct observations without influencing the teachers' behavior. Overall, 112 visits to directly trained took place. Third, we conducted *semi-structured interviews* with six participants in the Pedagogy group (about 120 min of audio recordings), six teachers in the Pedagogy & Content group (about 120 min), six peer teachers (about 70 min), and twelve government or TTU officials (about 150 min).[8]

3.3. Baseline characteristics, compliance, and attrition

Baseline characteristics are well-balanced across the three experimental groups (Table B.3). In particular, we report no significant differences in teachers' ($p = 0.65$) or students' ($p = 0.73$) standardized math scores. 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 3 out of 10 teachers in our sample are female, and the average teacher is 38 years old, graduated 12 years ago, and teaches 11 math lessons per week. Table B.18 shows substantial differences between target teachers and their peers: Target teachers are more likely to be male, score better in math, have more teaching experience in higher and less in lower grades, and are less likely to have only completed secondary education, suggesting that particularly strong candidates were selected for this role. Panel 2 on school characteristics shows that the typical class size is about 40 students.[9] The number of students who took the SFNA exam, roughly 50 per school, provides a proxy for the number of students per grade. Since this number is not much higher than the average class size, most schools can be assumed to have only one class per grade. About 92 percent of students in our sample passed the baseline math exam (with A, B, C or D), and 43 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 teachers in the treatment group participated in the five-day training course, and 94 percent of the teachers in the Pedagogy & Content group report using the laptops for content review. To be able to assess the impact of the program using students' test scores in grade 7, targeted teachers had to teach math to sixth graders in their school in 2020 and to seventh graders in 2021. Our endline teacher survey data show that 85 percent of the students in our sample were indeed taught by targeted teachers. This share does not differ significantly between experimental groups, suggesting that schools did not react strategically by assigning more math classes or different grades to treated teachers.

Tables B.4 and B.5 in the Online Appendix 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 (Table B.4). This yields a total sample size of 368 teachers. At the student level, we start with baseline data for 12,026 pupils from 220 schools. About 16 percent of these students either dropped out of school between grades 4 and 7, repeated grades, missed the endline examination, or could not be matched between the two rounds of testing, with dropout likely being the most common cause of attrition.[10] This leaves an estimation sample of 10,132 seventh graders from 220 schools. Both for teachers and students, attrition was unrelated to the experimental assignment (Table B.5). For students, this suggests that the program had no extensive margin effects on school dropout. For the estimation of spillovers, we

[7] We designed four different questionnaires: *(i)* a questionnaire for teachers in the Pedagogy group with items about the training and the implementation of the new methods, *(ii)* a similar questionnaire for the Pedagogy + Content group with additional questions about the content training with the laptops, *(iii)* a questionnaire for peer teachers asking about cascading activities, and *(iv)* a short questionnaire for the control group asking about the evaluation process. With the exception of the control group, the different survey versions followed the same basic structure and shared many common items, allowing for a comparison across different groups.

[8] During these conversations, the interviewees were asked to *(i)* share their general impression of the intervention, *(ii)* explain their views on the main elements of the Pedagogy intervention, *(iii)* share their views on the potential impact of the program on teachers' math and teaching skills as well as the learning outcomes of children, and *(iv)* give feedback on selected activities and program inputs. Additionally, officials were asked to *(v)* compare the pedagogical intervention with similar educational initiatives by other organizations, and *(vi)* comment on their attitudes toward rigorous program evaluation. Table E.1 in the Online Appendix provides an overview of statements by topic and type of interviewee.

[9] 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 fuzzy, since 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.

[10] We used name matching within schools to link endline observations from PSLE 2021 to the corresponding baseline observations from SFNA 2018. As 95 percent of the PSLE endline observations could be linked to their SFNA baseline (exact matches: 94%, fuzzy matches: 1%), attrition between the baseline and the endline assessments appears to be mainly due to school dropout or failure to take the PSLE examination. In addition, one expert teacher missed both data collection rounds. We impute the mean age, sex, and baseline score for this teacher in the student-level analysis.

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

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.113*	0.128**	0.046**	0.048**	0.023	0.029
	(0.065)	(0.064)	(0.023)	(0.023)	(0.023)	(0.022)
Pupil baseline score	0.473***	0.348***	0.117***	0.105***	0.204***	0.142***
	(0.017)	(0.024)	(0.007)	(0.009)	(0.008)	(0.010)
Control mean of dep. var.	0.000	0.000	0.124	0.124	0.570	0.570
Observations	10 132	10 132	10 132	10 132	10 132	10 132
Adjusted R^2	0.253	0.289	0.146	0.171	0.202	0.227
Controls (PDS Lasso)	No	Yes	No	Yes	No	Yes
Strata fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Lower Lee bound (90% CI)	−0.010	0.007	0.002	0.004	−0.019	−0.013
Upper Lee bound (90% CI)	0.257	0.269	0.092	0.094	0.071	0.075

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 score* is a pupil's math score in the SFNA exam administered in grade 4. Controls are selected using Post-Double Selection Lasso. Huber-White robust standard errors, clustered at the school level, in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

can use a sample of 15,073 grade 4 students from 220 schools. Due to the unavailability of student-level baseline data, we cannot study the attrition for this cohort of students.

4. Results

4.1. Did promoting participatory teaching strategies improve learning?

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

$$Y_{isk}^{PSLE(2021)} = \beta Treatment_s + \delta Y_i^{SFNA(2018)} + X_i'\gamma + V_s'\lambda + \phi_k + \epsilon_{isk}, \quad (1)$$

where $Y_{isk}^{PSLE(2021)}$ is the standardized math PSLE score of student i in school s and stratum k in 2021, and *Treatment* is a binary indicator that takes the value 1 if a school was assigned to receiving the pedagogical training (i.e., either Pedagogy or Pedagogy & Content) and is 0 otherwise.[11] All models control for students' math grades from the 2018 SFNA assessment, $Y_i^{SFNA(2018)}$, and strata fixed effects, ϕ_k. In our main specification, we additionally use Post-Double Selection Lasso (PDS Lasso) to select further controls (Belloni et al., 2014; Cilliers et al., 2024). The pool of potential controls includes 24 student-level variables, X_i, such as baseline performance in other subjects, and 24 school-level variables, V_s, such as geographical remoteness or teacher characteristics. Following Cilliers et al. (2024), we dummy out missings and include 27 missingness indicators as additional potential controls.[12] Standard errors are clustered at the school level in all student-level analyses.

Table 1 documents that students in treated schools significantly outperformed the control group by 0.11σ to 0.13σ (columns 1 and 2). Pupils in program schools were also 5 percentage points more likely to earn a top grade (i.e., A or B) than their peers in control schools (columns 3 and 4) corresponding to an increase in top grades by 37 to 39 percent. Estimates in columns 5 and 6 further suggest that the program increased pass rates by 2 to 3 percentage points, but these effects are not statistically significant at conventional levels. Student learning gains are very similar when separately estimated for the Pedagogy version (see Table 2), with 0.13σ to 0.14σ for overall scores, 6 percent for top grades, and a nonsignificant 2 to 3 percent for pass rates. This supports the finding that the pedagogical component of the teacher training drives the learning impact on students.

Fig. 2 and Table B.11 provide further insights into how the treatment affected the distribution of test scores. We observe shifts across all grades, with reductions in grades E, D, and C and increases in B and A. Effects are particularly pronounced for grade B, with average marginal effects of 4 percentage points (p = 0.029). Reductions in grade E (lowest possible grade) are also statistically significant, albeit small in magnitude due to the low baseline prevalence of this grade. Effects for other grades are not statistically significant individually.

Results are robust and similar across different specifications, with slightly increasing effect sizes when student, school, and teacher controls are included (Tables B.6 and B.7). We also estimate an ordered logit model for our main outcomes and obtain qualitatively similar results to the linear specification. This is consistent with evidence showing that linear regression is often robust to violations of interval-scale assumptions when applied to ordinal data (e.g., Norman, 2010). Table 1 also presents Lee (2009) bounds that account for minor variations in attrition rates (i.e., 15.5% in treatment vs 16.1% in control). The effects remain significant at the 10% level in models 2, 3, and 4 and extend slightly beyond zero in model 1.[13]

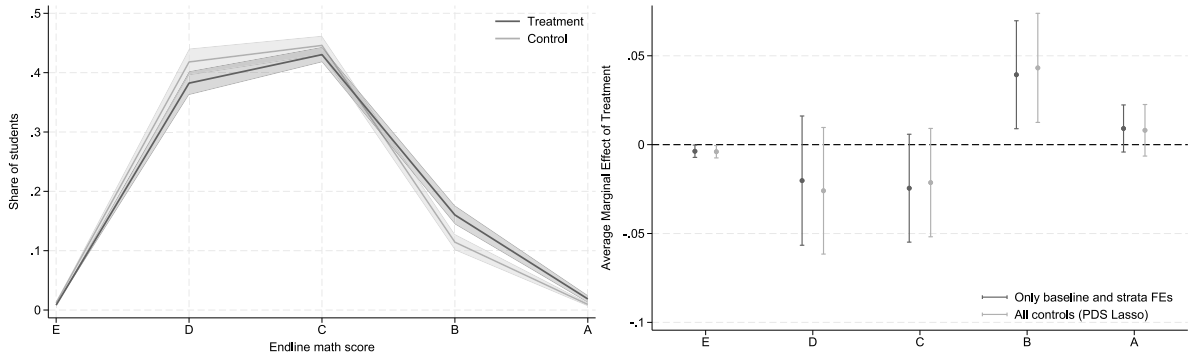
To contextualize effect sizes in educational interventions, Evans and Yuan (2022) reviewed 199 estimates from 75 randomized controlled trials focusing on math competency of students. Our estimated effect of 0.13σ places the participatory teaching program in the top 30 percent of educational interventions in terms of learning impact in math.

While our intervention targeted math teachers, these teachers teach an average of five subjects, including science (85%), social studies (79%), citizenship (68%), Swahili (57%), and English (48%) in addition to math (100%). Since the pedagogical strategies conveyed through the training—though typically illustrated with math examples—were not subject-specific, it is informative to analyze spillover effects on other subjects. Table B.10 in the Online Appendix examines effects

[11] In this first step, we pool both treatment versions, Pedagogy and Pedagogy & Content, for maximum precision before separately assessing the added value of CAL software in Section 4.2. This sequential approach maximizes statistical power when experimental data are limited while transparently presenting both treatment variants separately to address potential model selection concerns (Muralidharan et al., 2025).

[12] The models for our three outcome variables—i.e., standardized test scores, share scoring A or B, and pass rates—yield very similar specifications, with six variables, namely five student baseline scores and one missing-teacher-data indicator, being selected.

[13] Table 1 shows outer confidence intervals for Lee bounds rather than the Lee bounds themselves. Lee bounds for the main specifications (columns 2, 4, and 6) are [0.11, 0.14], [0.04, 0.05], and [0.03, 0.03] respectively. While confidence intervals span zero (column 1 in particular), the main specifications with controls show positive lower bounds, indicating our results are robust to worst-case assumptions about differential attrition. Moreover, if the differential attrition were systematic rather than random, it would likely bias our estimates downward. Student dropout is the main driver of attrition in our data and is concentrated among the worst-performing students. Since attrition is slightly lower in the treatment group, any systematic differential attrition would more likely understate rather than overstate our effect sizes.



(a) Distribution of endline math scores (b) Average Marginal Effects by grade

Fig. 2. Effects on the test score distribution and Average Marginal Effects on letter grades.
Notes: The left plot shows the distribution of endline math scores by treatment status. Shaded bands represent standard errors. The right plot presents average marginal effects from a multinomial logit model estimating the impact of the treatment on the probability of scoring in each outcome category (A to E, effects in percentage points). Whiskers indicate 90 percent confidence intervals. See Table B.11 for full regression results.

on students' average score across all subjects except math. Results for other subjects are similar but slightly less pronounced compared to math, with estimated effects of 0.09σ and an increase in top grades by up to 5 percentage points or 26 percent. This suggests that although the pedagogical training was tailored to math, teachers were able to transfer the methods to other subjects.

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; yet the absence of classroom observations in control schools limits our ability to assess whether these practices indeed represent changes from baseline teaching methods and differ from practices in control schools. As Figure D.1 in the appendix shows, treated teachers frequently employed methods such as group work (87% of visits), games (28%), student presentations (28%), and dialogs (26%). Treatment teachers also used a variety of instructional materials, including everyday objects (66% of visits), textbooks (46%), and flashcards (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 E.1 in the Online Appendix).

To better understand the circumstances under which the participatory teaching methods promoted by the training work best, it is instructive to look at how the effects vary by class, teacher, and student characteristics. A key challenge for productive student engagement is posed by the typically very large classes in Tanzania. According to Table B.8, the impact of the interventions decreases with larger class sizes, but these effects are only marginally significant (columns 7 and 8, see also Figures B.1a and B.2a). Another concern may be that the use of participatory teaching methods demands a higher level of content knowledge. Indeed, treatment effects appear to be larger for students who are taught by higher performing teachers (columns 5 and 6, see also Figures B.1b and B.2b). Additional analyses by students' sex and initial performance levels as well as teachers' sex, age, and experience 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

$$Y_{isk}^{PSLE(2021)} = \beta_1 T1_s + \beta_2 T2_s + \delta Y_i^{SFNA(2018)} + X_i'\gamma + V_s'\lambda + \phi_k + \epsilon_{isk}, \quad (2)$$

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. We use the same Post-Double Selection Lasso procedure as in Eq. (1).

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 Pedagogy & Content is slightly lower than the pure Pedagogy version, but this difference is very small and not significant ($\beta_2 - \beta_1 = -0.033$, p = 0.72).

To understand the mechanisms behind this null result for providing additional CAL content, it is informative to take a look at effects at the teacher level. Fig. 3 and Table B.13 present estimates of the causal effect of each treatment version on teachers' content knowledge in math. Although teachers who received laptops improved their understanding of concepts related to NSEA by 0.18σ (columns 5 and 6 in Table B.13), the effect on an overall score of math proficiency is smaller (0.13σ) and misses conventional levels of statistical significance (columns 1 and 2). Evidence from previous studies show that a 1σ increase in teacher content knowledge is associated with a 0.09σ improvement in student learning (Bau and Das, 2020; Metzler and Woessmann, 2012), suggesting that the effect on teachers was likely too small to translate into measurable differences at the student level.

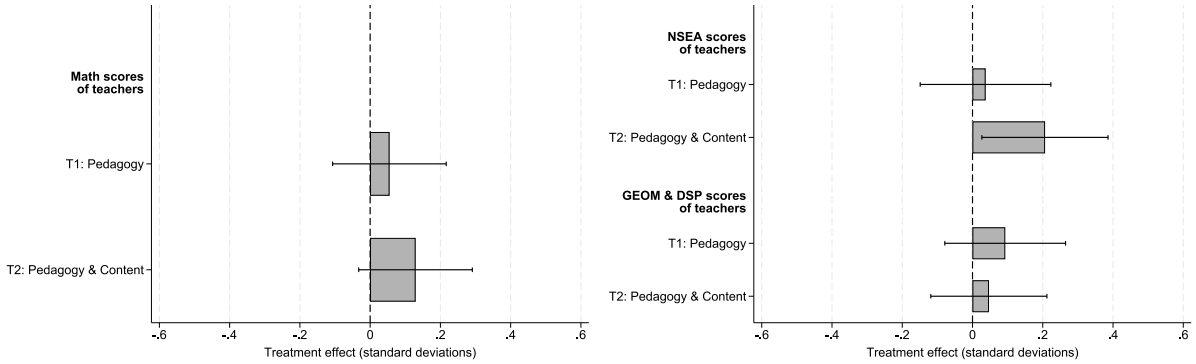
One possible interpretation of these modest effects is that teachers did not use or appreciate the laptops for their intended purpose. Our complementary data suggests otherwise. Teachers reported using the learning software for an average of 5 to 6 h weekly and gave very favorable evaluations of the computer-assisted learning aspect, with 68 percent rating it as excellent and 20 percent as good. This positive feedback was echoed in the interviews, where teachers unanimously voiced strong appreciation for the laptops and reported using them frequently for content revision or lesson preparation. As these self-reports may overstate actual use due to social desirability bias, we also inspect effect heterogeneity by age as a proxy for technology affinity. Indeed, we find some evidence that older teachers benefited less from the intervention, indicating potential technological barriers (Table B.17, columns 3 and 4). Another concern is that laptops may have crowded out lesson preparation time. We test this by comparing classroom observations from treatment teachers with and without laptops. We find no evidence of crowding-out: lesson preparation quality (3.88 vs 3.76 out of 5, p = 0.34), on-time class starts (>95% in both groups), and share of children on-task (35% vs 33%, p = 0.82) were similar across groups. Of six teaching elements observed, only "dialog" differed significantly between groups (p = 0.06).

Another plausible explanation is that most teachers already had a good command of the primary school curriculum to begin with. As

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	0.134	0.140*	0.057*	0.057**	0.024	0.027
	(0.085)	(0.081)	(0.029)	(0.029)	(0.028)	(0.027)
T2: Pedagogy & Content	0.091	0.107	0.035	0.038	0.022	0.028
	(0.076)	(0.074)	(0.026)	(0.026)	(0.029)	(0.028)
Pupil baseline math score	0.473***	0.349***	0.117***	0.105***	0.204***	0.142***
	(0.017)	(0.024)	(0.007)	(0.009)	(0.008)	(0.010)
T2 − T1	−0.043	−0.033	−0.022	−0.019	−0.002	0.001
	(0.094)	(0.092)	(0.033)	(0.032)	(0.033)	(0.032)
Control mean of dep. var.	0.000	0.000	0.124	0.124	0.570	0.570
Observations	10 132	10 132	10 132	10 132	10 132	10 132
Adjusted R^2	0.253	0.288	0.147	0.171	0.202	0.227
Controls (PDS Lasso)	No	Yes	No	Yes	No	Yes
Strata 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 are selected using Post-Double Selection Lasso. Huber-White robust standard errors, clustered at the school level, in parentheses. * p<0.10, ** p<0.05, *** p<0.01.



(a) Effects on overall test score (b) Effect on test scores by domain

Fig. 3. Treatment effects on teachers' overall and domain-specific math scores *Notes:* 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 the Online Appendix B.7.

shown in Figure B.4, the average teacher was able to correctly 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 high-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). Hence, it appears plausible that many Tanzanian teachers in our sample regions are already sufficiently proficient in math to teach effectively at the primary level. In line with this argument, Table B.17 in the Online Appendix points to considerable effect heterogeneity by teachers' initial ability level. Low-performing teachers markedly improved their content knowledge as a result of the intervention (0.51σ, p = 0.004, for teachers below the median), but these effects decline significantly as teachers' baseline scores improve, and are close to zero for high-performing teachers (not shown). However, this improvement in low-performing teachers' content knowledge did not translate into gains in student learning (Figure B.1b).

From an impact evaluation perspective, the additional investment in the IT equipment for content revision did not pay off. Although low-performing teachers appear to have used the software to catch up with their better-prepared colleagues, we do not find that such gains were transferred to students. This is also in line with recent studies reporting mixed evidence on the effectiveness of education technology (e.g., Beg et al., 2022; de Barros, 2023).

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

To analyze spillovers on indirectly exposed fourth graders rather than directly exposed seventh graders, we slightly adapt Equation (1) and estimate

$$Y_{isk}^{SFNA(2021)} = \beta Treatment_s + \delta \bar{Y}_s^{SFNA(2018)} + X_i'\gamma + V_s'\lambda + \phi_k + \epsilon_{isk}, \quad (3)$$

where $Y_{isk}^{SFNA(2021)}$ is the standardized math SFNA score of student i in school s and stratum k in 2021. As no standardized national assessment results are published for students below grade four, we include the school-level SFNA score from 2018, $\bar{Y}_s^{SFNA(2018)}$, as a lagged performance measure. Our set of potential controls for the Post-Double Selection Lasso includes student gender, the school-level variables from Eq. (1) and various school-level aggregates of student-level characteristics.[14]

[14] Depending on the outcome, this procedure selects one to three additional controls, including school-level baseline scores, student gender, and the pupil–teacher ratio.

Table 3
Cascading effect on students' math scores.

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.029	0.049	0.011	0.011	−0.000	0.014
	(0.051)	(0.049)	(0.010)	(0.010)	(0.018)	(0.016)
School SFNA avg. score (std)	0.155***	0.125***	0.025***	0.026***	0.055***	0.041***
	(0.032)	(0.033)	(0.006)	(0.006)	(0.012)	(0.011)
Control mean of dep. var.	0.000	0.000	0.066	0.066	0.730	0.730
Observations	15 073	15 073	15 073	15 073	15 073	15 073
Adjusted R^2	0.068	0.075	0.032	0.036	0.041	0.055
Controls (PDS Lasso)	No	Yes	No	Yes	No	Yes
Strata 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 pre-treatment scores in the SFNA exam administered in grade 4 and the PSLE exam administered in grade 7. Controls are selected using Post-Double Selection Lasso. Huber-White robust standard errors, clustered at the school level, in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 3 examines spillover effects on students whose teachers were indirectly exposed to the treatment through peer learning activities in their school. In all pooled specifications, estimates are small and not statistically significant. When disaggregating by program version in Table B.9, the pure Pedagogy treatment yields larger point estimates, though they are statistically insignificant in five out of six specifications. At the teacher level, we find mixed evidence of content knowledge spillovers. For the Pedagogy & Content treatment, we estimate a statistically insignificant spillover effect of 0.14σ on peer teachers (Tables B.13 and B.15), which is consistent with the moderate direct effects of the additional content training. In contrast, the Pedagogy arm yields a significant increase of 0.21σ in peer teacher math competency ($p = 0.05$, Table B.15).[15]

One possible explanation for the absence of compelling treatment externalities is that the observation period of our study was not long enough to capture effects on students of indirectly exposed teachers. Given 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. Another concern is that fourth and seventh grades differ in teacher quality, class size, and curriculum complexity, and those differences may influence treatment effectiveness regardless of cascading (see Table B.8 for the heterogeneity analysis). Relatedly, the distributional properties of the grade-specific assessments—like potential ceiling or floor effects—vary between fourth and seventh grades (see Figures 2(a) and B.3a).[16] To assess the plausibility of these hypotheses, 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 the PSLE scores for these years, we conduct a difference-in-difference analysis to assess the impact of the program for seventh graders over a longer time horizon. Unlike our 2020/21 experimental evaluation, previous implementation rounds did not target specific grade levels. Typically, one teacher per school participated in the training, with all other teachers exposed only indirectly through cascading activities. Hence, our difference-in-difference estimates correspond to an upper bound for spillover effects at the school level. As Fig. 4 and extensive analyses in Online Appendix C show, we obtain a null effect for this upper bound estimate of spillovers. Hence, neither insufficient time for peer teachers to adopt new techniques nor variation between fourth grade and seventh grade (e.g., teacher quality, distributional properties of assessments) seem plausible explanations for the lack of robust spillovers.

Another possibility is that the knowledge sharing activities were not carried out. 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.[17]

A further consideration is that differential attrition among fourth graders might bias our spillover estimates downward. If student dropout is concentrated among the worst-performing fourth graders and the professional development program reduces dropout, we would underestimate the spillover effect. While we cannot directly assess this potential confounding channel, it seems unlikely given the minimal variation in attrition rates among seventh graders (Table B.5).

Hence, a more plausible 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 cost-effective is the program?

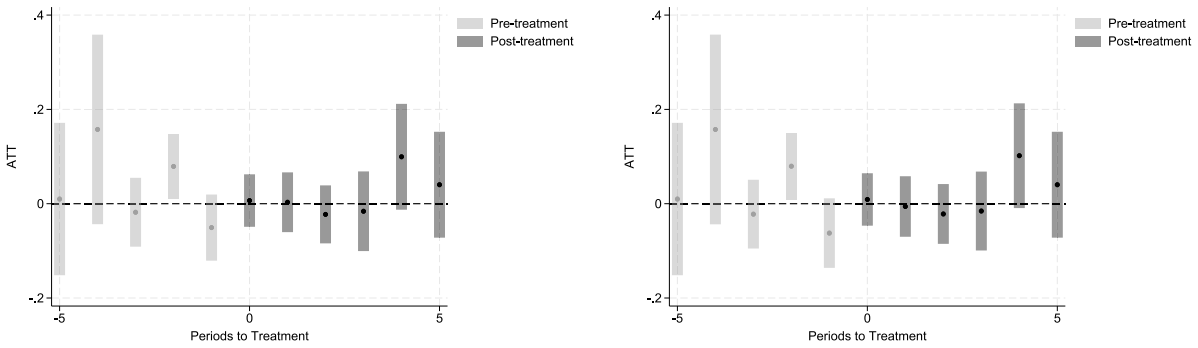
Due to potential externalities and second-round effects, accurately assessing the cost-effectiveness of teacher professional development is challenging. While our design captures spillovers to other teachers, it cannot account for gains to future student cohorts taught by trained teachers. Thus, our estimates likely represent a lower bound for the true cost-effectiveness.

With this limitation in mind, the implementation costs for the training program amount to 760 USD per teacher and 14 USD per student. The main cost items were staff salaries and allowances for teachers and government officials participating in the intervention. In the program arm combining pedagogy and subject-specific content, the procurement of tablets equipped with CAL software raised costs by 41 percent, to 1070 USD per teacher. Based on these estimates, every

[15] This finding allows for two interpretations: *First*, target teachers—who had higher baseline math skills (Table B.18)—may have helped peer teachers refresh knowledge during knowledge sharing activities, even though they themselves showed no further gains (Table B.13). *Second*, the result might reflect random variation, particularly given that the Pedagogy & Content treatment—which explicitly included content training—produced no measurable spillover effects on peer teachers.

[16] While PSLE scores roughly follow a bell-shaped curve, SFNA scores are more compressed at the lower end, with nearly 30 percent of students scoring an E (lowest grade), implying less sensitivity to potential improvements among low-performing students. This difference is unlikely to account for the null results for spillovers, as direct effects were mainly driven by changes at the upper end of the distribution.

[17] This should not be taken as conclusive evidence of the successful implementation of the cascading elements, as teachers may have succumbed to a common tendency to give socially desirable, but dishonest answers. Indeed, in the in-depth interviews, teachers were somewhat more critical of the cascading elements, with some interviewees mentioning challenges to their implementation due to a lack of interest from some of their colleagues.



(a) Never treated (b) Not yet treated

Fig. 4. Event study plots for cascading effects, 2013–2019. *Notes:* Event study estimates based on Callaway and Sant'Anna (2021) for PSLE (grade 7) test scores. Estimates compare treated schools to never treated schools (A) or not-yet-treated schools (B). Period 0 represents the year of treatment. 95% confidence intervals shown.

100 USD invested in the teacher training program increased students' math scores by 1.0σ. For the Pedagogy & Content arm, cost-effectiveness decreases to 0.5σ per 100 USD, as the tablets raised program costs without improving impact (Table 2).

Even though we do not report statistically significant spillover effects through peer-to-peer cascading, it is illustrative to gauge how such effects would impact the cost-effectiveness. With indirectly exposed students outnumbering those of trained teachers by eight to one, even small spillovers could increase program benefits several-fold. If we use our statistically insignificant point estimate of 0.042σ as the true spillover effect (Table 3), cost-effectiveness would increase from 1.0σ per 100 USD to 3.4σ per 100 USD.

To contextualize our cost-effectiveness estimates, we compare them to education programs reviewed by Kremer et al. (2013) and Angrist et al. (2025). To enhance comparability to the studies reviewed in Kremer et al. (2013), we adjust our main cost-effectiveness estimate of 1.0σ per 100 USD^{2020} for inflation between 2011 and 2020, that is, between the base year used in the review and the implementation year of our study. This yields a cost-effectiveness estimate of 1.2σ per 100 USD^{2011}. Among the 30 interventions reviewed by Kremer et al. (2013), 16 (53%) report a negative or zero effect, 6 (20%) show higher cost-effectiveness, and 8 (27%) interventions estimate a similar cost-effectiveness, with 1.2σ per 100 USD^{2011} falling within their 90% confidence interval. Angrist et al. (2025) synthesize cost-effectiveness estimates (without inflation or PPP adjustments) for 97 studies with 135 different treatment arms. For those 135 interventions, 48 (36%) did not produce statistically significant effects, the median cost-effectiveness is 0.39σ, and a cost-effectiveness of 1.0σ per 100 USD corresponds to the 62th percentile.

In summary, the results suggest that, even based on conservative estimates excluding effects on future student cohorts or spillovers from cascading, pedagogical teacher training can be a cost-effective approach to improving student learning.

4.5. 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 more qualitative tradition focusing on the experiences of project stakeholders (e.g., Banerjee and Duflo, 2009; Garbarino and Holland, 2009). Despite the emphasis on causal inference methods within academic circles, many practitioners favor experience-based approaches, and few programs undergo rigorous impact evaluations. The main contribution of this paper is the identification of causal effects through a field experiment, but we can also combine and compare our experimental findings with insights from surveys and interviews with project beneficiaries. In particular, we asked all participating teachers to gauge 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).

These findings tie into a nascent literature studying biases in evaluations (e.g., Camfield et al., 2014). Two broad explanations accounting for participants' overly optimistic 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 McKenzie (e.g., 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 feedback gathered through participant surveys and interviews, while a valuable complement to experimental evidence, are ill-suited for assessing causal effects.

5. Conclusion

Addressing the 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). Teachers are critical to the success of an education system, make up a substantial share of the global workforce and account for 80 percent of educational expenditures (UNESCO, United Nations Educational, Scientific and Cultural Organization, 2024). 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. In 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 0.13σ. Our findings are based on standardized national assessments 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

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

	RCT: Causal estimates	Survey: Participants' beliefs about impact
Impact of intervention on student learning	Significant effect of 0.13 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 0.13 SD, but insignificant	Did the project improve your math skills? Strongly agree: 85%, rather agree: 11%
Spillovers of Pedagogy & Content *intervention on peer teachers' math skills*	Effect of 0.14 SD, but insignificant	Did the project improve your math skills? Strongly agree: 81%, rather agree: 15%

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 in student learning would require a 1.1σ improvement in teachers' content knowledge (Bau and Das, 2020; Metzler and Woessmann, 2012)—a very ambitious effect size 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 to 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 the reach of a project, yet our results suggest that producing measurable learning spillovers is not straightforward. More research is 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 cost-effective approach to improve student learning. Teachers influence dozens of student cohorts throughout their careers, and evidence suggests that improved teaching practices can persist across multiple cohorts, even if effects may diminish over time (e.g., Cilliers et al., 2022a). Promoting participatory teaching may thus be a key ingredient in a comprehensive strategy to ensure that children in developing countries are not only going to school, but actually learning.

CRediT authorship contribution statement

Martina Jakob: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Konstantin Büchel:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Daniel Steffen:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Aymo Brunetti:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jdeveco.2026.103742.

Data availability

The anonymized data is available in the Harvard Dataverse, DOI: https://doi.org/10.7910/DVN/DH3PZA.

References

Angrist, Noam, Evans, David, Filmer, Deon, Glennerster, Rachel, Rogers, Halsey, Sabarwal, Shwetlena, 2025. How to improve education outcomes most efficiently? A review of the evidence using a unified metric. J. Dev. Econ. 172, 103382, URL https://www.sciencedirect.com/science/article/pii/S0304387824001317.

Banerjee, Abhijit, Banerji, Rukmini, Berry, James, Duflo, Esther, Kannan, Harini, Mukerji, Shobhini, Shotland, Marc, Walton, Michael, 2017. From proof of concept to scalable policies: Challenges and solutions, with an application. J. Econ. Perspect. 31 (4), 73–102.

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

Barrett, Allison, 2010. Monitoring and evaluating INSET in India: Challenges and possible solutions. In: Sheehan, Susan (Ed.), Teacher Development and Education in Context. British Council, pp. 5–16.

Bau, Natalie, Das, Jishnu, 2020. Teacher value-added in a low-income country. Am. Econ. J.: Econ. Policy 12 (1), 62–96.

Beg, Sabrin, Halim, Waqas, Lucas, Adrienne M., Saif, Umar, 2022. Engaging teachers with technology increased achievement, bypassing teachers did not. Am. Econ. J.: Econ. Policy 14 (2), 61–90.

Belloni, Alexandre, Chernozhukov, Victor, Hansen, Christian, 2014. Inference on treatment effects after selection among high-dimensional controls. Rev. Econ. Stud. 81 (2), 608–650.

Berlinski, Samuel, Busso, Matias, 2017. Challenges in educational reform: An experiment on active learning in mathematics. Econ. Lett. 156, 172–175.

Bett, Harry Kipkemoi, 2016. The cascade model of teachers continuing professional development in Kenya: A time for change? Cogent Educ. 3 (1), 1–9.

Bietenbeck, Jan, Piopiunik, Marc, Wiederhold, Simon, 2018. Africa's skill tragedy: Does teachers' lack of knowledge lead to low student performance? J. Hum. Resour. 53 (3), 553–578.

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

British Council, 2018. Train the trainers and cascade models: A practical guide and toolkit. Published online, https://www.britishcouncil.org/education/skills-employability/tool-resources/train-trainers (Accessed: 01 November 2024).

Brunetti, Aymo, Büchel, Konstantin, Jakob, Martina, Jann, Ben, Kühnhanss, Christoph, Steffen, Daniel, 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, Büchel, Konstantin, Jakob, Martina, Jann, Ben, Steffen, Daniel, 2023. Inadequate teacher content knowledge and what could be done about it: Evidence from El Salvador. J. Dev. Eff. 16 (2), 1–24.

Büchel, Konstantin, Jakob, Martina, Christoph, Kühnhanss, Steffen, Daniel, Brunetti, Aymo, 2022. The relative effectiveness of teachers and learning software. Evidence from a field experiment in El Salvador. J. Labor Econ. 40 (3), 737–777.

Buhl-Wiggers, Julie, Kerwin, Jason, de la Piedra, Ricardo Montero, Smith, Jeffrey, Thornton, Rebecca, 2023. Reading for life: Lasting impacts of a literacy intervention in Uganda. Mimeo.

Callaway, Brantly, Sant'Anna, Pedro, 2021. Difference-in-differences with multiple time periods. J. Econometrics 225 (2), 200–230.

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

Christensen, Anders Astrup, Jerrim, John, 2025. Professional learning communities and teacher outcomes: A cross-national analysis. Teach. Teach. Educ. 156, 104920.

Cilliers, Jacobus, Elashmawy, Nour, McKenzie, David, 2024. Using Post-Double Selection Lasso in Field Experiments. World Bank.

Cilliers, Jacobus, Fleisch, Brahm, Kotze, Janeli, Mohohlwane, Mpumi, Taylor, Stephen, 2022a. The challenge of sustaining effective teaching: Spillovers, fade-out, and the cost-effectiveness of teacher development programs. Econ. Educ. Rev. 87, 102215.

Cilliers, Jacobus, Fleisch, Brahm, Kotze, Janeli, Mohohlwane, Nompumelelo, Taylor, Stephen, Thulare, Tsegofatso, 2022b. Can virtual replace in-person coaching? Experimental evidence on teacher professional development and student learning. J. Dev. Econ. 155, 102815.

Cilliers, Jacobus, Fleisch, Brahm, Prinsloo, Cas, Taylor, Stephen, 2020. How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. J. Hum. Resour. 55 (3), 926–962.

Cilliers, Jacobus, Habyarimana, James, 2023. Tackling implementation challenges with information: Experimental evidence from a school governance reform in tanzania. RISE Working Paper 23/142.

Cilliers, Jacobus, Mbiti, Isaac M., Zeitlin, Andrew, 2021. Can public rankings improve school performance? Evidence from a nationwide reform in Tanzania. J. Hum. Resour. 56 (3), 655–685.

Cornelius-White, Jeffrey, 2007. Learner-centered teacher-student relationships are effective: A meta-analysis. Rev. Educ. Res. 77 (1), 113–143.

de Barros, Andreas, 2023. Explaining the Productivity Paradox: Experimental Evidence from Educational Technology. ERIC, Annenberg Institute for School Reform at Brown University, EdWorkingPaper No. 23-853.

de Barros, Andreas, Henry, Junita, Mathenge, Jacqueline, 2021. What drives teachers to change their instruction? A mixed-methods study from zambia. Unpublished manuscript. Cambridge, MA.

De Ree, Joppe, Muralidharan, Karthik, Pradhan, Menno, Rogers, Halsey, 2018. Double for nothing? Experimental evidence on an unconditional teacher salary increase in Indonesia. Q. J. Econ. 133 (2), 993–1039.

Dichaba, Mpho, Mokhele, Matseliso, 2012. Does the cascade model work for teacher training? Analysis of teachers experiences. Int. J. Educ. Sci. 4 (3), 249–254.

Duflo, Esther, Hanna, Rema, Ryan, Stephen P., 2012. Incentives work: Getting teachers to come to school. Am. Econ. Rev. 102 (4), 1241–1278.

Escueta, Maya, Nickow, Andre, Oreopoulos, Philip, Quant, Vincent, 2020. Upgrading education with technology: Insights from experimental research. J. Econ. Lit. 58 (4), 897–996.

Evans, David K., Mendez Acosta, Amina, 2021. Education in Africa: What are we learning? J. Afr. Econ. 30 (1), 13–54.

Evans, David K., Popova, Anna, 2016. What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. World Bank Res. Obs. 31 (2), 242–270.

Evans, David K., Yuan, Fei, 2022. How big are effect sizes in international education studies? Educ. Eval. Policy Anal. 44 (3), 532–540.

Ganimian, Alejandro J., Murnane, Richard J., 2016. Improving education in developing countries: Lessons from rigorous impact evaluations. Rev. Educ. Res. 86 (3), 719–755.

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

GEEAP, Global Education Evidence Advisory Panel, 2020. Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are Smart Buys for Improving Learning in Low- and Middle-Income Countries?. World Bank, Recomendations published by the World Bank, the FCDO, and Building Evidence in Education, October 2020.

Glewwe, Paul, Muralidharan, Karthik, 2016. Improving education outcomes in developing countries: Evidence, knowledge gaps and policy implications. In: Hanushek, Eric, Machin, Stephen, Woessmann, Ludger (Eds.), Handbook of the Economics of Education. Elsevier, Amsterdam, pp. 653–743.

Gore, Jennifer M., Miller, Andrew, Fray, Leanne, Harris, Jess, Prieto, Elena, 2021. Improving student achievement through professional development: Results from a randomised controlled trial of quality teaching rounds. Teach. Teach. Educ. 101, 103297.

Gray-Lobe, Guthrie, Keats, Anthony, Kremer, Michael, Mbiti, Isaac, Ozier, Owen, 2022. Can education be standardized? Evidence from Kenya. University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2022-68.

Hale, Thomas, Angrist, Noam, Goldszmidt, Rafael, Kira, Beatriz, Petherick, Anna, Phillips, Toby, Webster, Samuel, Cameron-Blake, Emily, Hallas, Laura, Majumdar, Saptarshi, Tatlow, Helen, 2021. A global panel database of pandemic policies. Nat. Hum. Behav. 5, 529–538.

Harbour, Kristin E., Evanovich, Lauren L., Sweigart, Chris A., Hughes, Lindsay E., 2015. A brief review of effective teaching practices that maximize student engagement. Prev. Sch. Fail. 59 (1), 5–13.

Jacob, Brian A., 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. J. Public Econ. 89 (5–6), 761–796.

Jacob, Andy, McGovern, Kate, 2015. The mirage: Confronting the hard truth about our quest for teacher development. The New Teacher Project, published online: https://eric.ed.gov (Access: 28 December 2024).

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

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

Kremer, Michael, Brannen, Conner, Glennerster, Rachel, 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. Qual. Quant. 47 (4), 2025–2047.

Lee, David S., 2009. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Rev. Econ. Stud. 76 (3), 1071–1102.

Loyalka, Prashant, Popova, Anna, Li, Guirong, Shi, Zhaolei, 2019. Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program. Am. Econ. J.: Appl. Econ. 11 (3), 128–154.

Marinelli, Horacio Alvarez, Berlinski, Samuel, Busso, Matías, Correa, Julián Martínez, 2023. Improving early literacy through teacher professional development: Experimental evidence from Colombia. J. Public Econ. Plus 4, 100019.

Mbiti, Isaac, Muralidharan, Karthik, Romero, Mauricio, Schipper, Youdi, Manda, Constantine, Rajani, Rakesh, 2019. Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. Q. J. Econ. 134 (3), 1627–1673.

Mbiti, Isaac, Romero, Mauricio, Schipper, Youdi, 2023. Designing effective teacher performance pay programs: Experimental evidence from Tanzania. Econ. J. 133 (653), 1968–2000.

McKenzie, David, 2018. Can business owners form accurate counterfactuals? Eliciting treatment and control beliefs about their outcomes in the alternative treatment status. J. Bus. Econom. Statist. 36 (4), 714–722.

Metzler, Johannes, Woessmann, Ludger, 2012. The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. J. Dev. Econ. 99 (2), 486–496.

Miguel, Edward, Kremer, Michael, 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. Econometrica 72 (1), 159–217.

Muralidharan, Karthik, Romero, Mauricio, Wüthrich, Kaspar, 2025. Factorial designs, model selection, and (incorrect) inference in randomized experiments. Rev. Econ. Stat. 107 (3), 589–604.

Muralidharan, Karthik, Sundararaman, Venkatesh, 2011. Teacher performance pay: Experimental evidence from India. J. Political Econ. 119 (1), 39–77.

NECTA, 2018. Format for Standard Four National Assessment. Tech. Rep., National Examinations Council of Tanzania.

NECTA, 2020. Format for Primary School Leaving Examinations. Tech. Rep., National Examinations Council of Tanzania.

Norman, Geoff, 2010. Likert scales, levels of measurement and the "laws" of statistics. Adv. Health Sci. Educ. 15 (5), 625–632.

Nourani, Vesall, Ashraf, Nava, Banerjee, Abhijit, 2023. Learning to teach by learning to learn. Mimeo.

OECD, Organisation for Economic Co-operation and Development, 2019. TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners. OECD Publishing, Paris.

Orr, David, Westbrook, Jo, Pryor, John, Durrani, Naureen, Sebba, Judy, 2013. What are the Impacts and Cost-Effectiveness of Strategies to Improve Performance of Untrained and Under-Trained Teachers in the Classroom in Developing Countries?. EPPI-Centre, Institute of Education, University of London, London.

Patrinos, Harry A., Angrist, Noam, 2018. Global dataset on education quality: A review and update (2000–2017). World Bank Policy Research Working Paper No. 8592.

PCA, Parliamentary Control of the Administration, 2023. Report of the parliamentary control of the administration for the attention of the council of states control committee, 27 april 2023.

Piper, Benjamin, Sitabkhan, Yasmin, Mejia, Jessica, Betts, Kellie, 2018. Effectiveness of teachers' guides in the global south: Scripting, learning outcomes, and classroom utilization. RTI Int..

Popova, Anna, Evans, David, Breeding, Mary, Arancibia, Violeta, 2022. Teacher professional development around the world: The gap between evidence and practice. World Bank Res. Obs. 37 (1), 107–136.

Romero, Mauricio, Bedoya, Juan, Yanez-Pagans, Monica, Silveyra, Marcela, De Hoyos, Rafael, 2022. Direct vs indirect management training: Experimental evidence from schools in Mexico. J. Dev. Econ. 154, 102779.

SDC, Swiss Agency for Development and Cooperation, 2024. Evaluation reports. Overview of evaluations and reports from the specialist service Evaluation and Corporate Controlling in the database. https://www.eda.admin.ch/deza/en/home/results-impact/berichte/evaluationsberichte.html. (Accessed: 28 December 2024).

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

Singh, Abhijeet, 2024. Improving administrative data at scale: Experimental evidence on digital testing in Indian schools. Econ. J. 134 (661), 2207–2223.

Smith, Jeffrey, Whalley, Alexander, Wilcox, Nathaniel, 2021. Are Participants Good Evaluators?. W.E. Upjohn Institute for Employment Research, Michigan.

Sumra, Suleman, Ruto, Sara, Rajani, Rakesh, 2015. Assessing literacy and numeracy in Tanzania's primary schools: The uwezo approach. In: Joshi, A.R., Gaddis, I. (Eds.), Preparing the Next Generation in Tanzania. World Bank Group, Washington D.C., pp. 47–64.

UN, United Nations, 2015. The 2030 Agenda for Sustainable Development. United Nations, New York.

UNESCO, United Nations Educational, Scientific and Cultural Organization, 2020. Primary school pupil-teacher ratio, tanzania. Published online: https://data.worldbank.org/indicator/SE.PRM.ENRL.TC.ZS?locations=TZ. (Accessed: 30 December 2024).

UNESCO, United Nations Educational, Scientific and Cultural Organization, 2024. Staff compensation as share of expenditure in primary public institutions. Processed by Our World in Data and published online. (Accessed: 30 December 2024).

UNICEF, 2024. Data Must Speak: Unpacking Factors Influencing School Performance in Mainland Tanzania. UNICEF, Florence, Italy, Innocenti research report, URL https://www.unicef.org/innocenti/media/5696/file/UNICEF-Innocenti-DMS-Mainland-Tanzania-Report-2024.pdf. (Accessed 27 June 2025).

USAID, United States Agency for International Development, 2024. Evaluations at USAID - Dashboard. https://www.usaid.gov/evaluation/evaluations-usaid-dashboard. (Accessed: 26 June 2024).

Vaughan, Tanya, Richardson, Sarah, Carslake, Toby, Reimers, Trisha, Macaskill, Greg, Newton, Toby, Zoanetti, Nathan, Mannion, Andrew, Murphy, Martin, 2023. Building capacity for quality teaching rounds–Victoria. Final report by the University of Newcastle. Online available, URL: https://doi.org/10.37517/978-1-74286-713-7.

Vescio, Vicki, Ross, Dorene, Adams, Alyson, 2008. A review of research on the impact of professional learning communities on teaching practice and student learning. Teach. Teach. Educ. 24 (1), 80–91.

Wolf, Sharon, Aber, Lawrence, Behrman, Jere, Tsinigo, Edward, 2019. Experimental impacts of the 'quality preschool for Ghana' interventions on teacher professional well-being, classroom quality, and children's school readiness. J. Res. Educ. Eff. 12 (1), 10–37.

World Bank, 2018. World Development Report 2018: Learning to Realize Education's Promise. World Bank, 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/. (Accessed: 26 June 2024).

Zizzo, Daniel John, 2010. Experimenter demand effects in economic experiments. Exp. Econ. 13, 75–98.