JICA Ogata Research Institute Discussion Paper

Can Teacher Training Improve Teacher Competence and Student Learning? Evidence from Ethiopia

Eiji Kozuka

No. 43
October 2025

JICA OGATA SADAKO RESEARCH INSTITUTE FOR PEACE AND DEVELOPMENT

Suggested Citation: Kozuka E. 2025. Can Teacher Training Improve Teacher Competence and Student Learning? Evidence from Ethiopia. JICA Ogata Research Institute Discussion Paper No.43. Tokyo: JICA Ogata Research Institute for Peace and Development.

JICA Ogata Sadako Research Institute for Peace and Development, Japan International Cooperation Agency (JICA)
10-5 Ichigaya Honmura-cho, Shinjuku-ku, Tokyo, 162-8433, JAPAN
TEL: +81-3-3269-3374
FAX: +81-3-3269-2054

Can Teacher Training Improve Teacher Competence and Student Learning? Evidence from Ethiopia

Eiji Kozuka*

Abstract

Teacher quality is a critical factor in student learning. However, little is known about how low- and middle-income countries can enhance teacher quality within their government systems. This study presents results from a randomized evaluation of a government-implemented teacher professional development (PD) program for primary school mathematics and physics teachers in Ethiopia. As a complementary intervention to teacher training, this study also tested the effectiveness of using workbooks to help students retain what they learned in class. Following teacher training, teachers improved their teaching practices and content knowledge, but their students did not demonstrate short-term improvements in test scores without the use of workbooks. When students used workbooks with teacher assistance in supplemental classes, however, their test scores improved. After these interventions, teachers continued their efforts in their schools without financial incentives, reinforcing the impacts on teaching practice and teacher knowledge.

Keywords: Education policy, Student learning, Teacher training, Professional development

* Professor, School of International Relations, University of Shizuoka (kozuka.eiji@u-shizuoka-ken.ac.jp)

I am deeply grateful to Satoshi Kusaka, Nobuko Chiku, Akane Morimoto, Mizuki Matsuzaki, Shimpei Taguchi, and the officials of the Oromia Regional Education Bureau and JICA for their assistance in implementing the intervention. I also thank Eriko Yagi, Mulu Teka, and his team for their support in collecting and managing the data. I appreciate the invaluable comments and advice from Takashi Kurosaki, Yukichi Mano, Atsushi Matachi, Norihiro Nishikata, Takeshi Aida, and seminar participants at JICA and Hitotsubashi University.

1. Introduction

The need to improve teacher quality is growing rapidly in low- and middle-income countries. Although teachers play a key role in improving student learning (Rivkin, Hanushek, and Kain 1998; Rockoff 2004; Hanushek and Rivkin 2010; Chetty, Friedman, and Rockoff 2014), many teachers in these countries lack the necessary pedagogical skills and content knowledge. A study of primary school teachers in seven countries in sub-Saharan Africa found that only 68 percent of the teachers surveyed had minimum content knowledge in numeracy, and only 11 percent had minimum general pedagogical knowledge, which was reflected in poor classroom practice (Bold et al. 2017). [1]

To improve the skills and knowledge of teachers, countries around the world allocate substantial budgets toward implementing teacher professional development (PD) programs. However, few studies have rigorously evaluated the effectiveness of teacher PD programs in low- and middle-income countries. The question of how to design effective teacher PD programs remains unresolved for policymakers and practitioners (Snilstveit et al. 2015; Evans and Popova 2016; Popova et al. 2022).

To address this knowledge gap, this study conducted a randomized experiment in 335 schools in Ethiopia to evaluate a teacher PD program for grade seven and eight mathematics and physics teachers, as well as mathematics and physics workbooks for students to complement the PD program. The program was implemented from 2015 to 2016 by the Federal Ministry of Education in Ethiopia and the Oromia Regional Education Bureau with technical assistance from the Japan International Cooperation Agency (JICA).

In this experiment, two treatment groups were created: (i) the first treatment group (referred to as Regular PD) received in-service training for mathematics and physics teachers, and (ii) the second treatment group (referred to as PD-plus-Workbook) received student workbooks in mathematics and physics as supplementary materials to help students retain what they learned in class by completing exercises. The workbook content was designed to align with the PD content, and the students used the workbooks with teacher assistance in supplemental classes. In addition to the same in-service training as the Regular PD group, workshops were conducted for school principals and teachers on how to effectively utilize the workbooks. Participants were not provided with financial incentives or incentives directly tied to promotion during and after the interventions, which are frequently offered by other donor-funded programs. To evaluate the impacts of these interventions on teacher quality and student learning, we collected data on teaching practices, teachers' content knowledge, and student test scores.

---

[1] In this study, teachers are expected to possess minimum subject content knowledge in numeracy if they can correct children's work with at least 80 percent accuracy, and minimum general pedagogy knowledge if they score at least 80 percent on specific pedagogical tasks (Bold et al. 2017).

Six months after the Regular PD program for teachers, we found that both the Regular PD group and the PD-plus-Workbook group teachers had improved their teaching practices and content knowledge. Furthermore, ten months after the program, we observed that teachers continued their PD practices at their schools, further enhancing the impact on their teaching and knowledge.

Regarding student test scores, although Regular PD alone did not lead to improvements, the PD-plus-Workbook approach produced measurable gains. This outcome was not surprising, as the PD was not intended to directly boost student test scores, whereas the workbooks were designed to help students retain what they learned in class—crucial for improving those scores.

This study adds to the literature on improving teacher effectiveness and student learning in low- and middle-income countries. Based on a growing number of experimental studies in this area, three types of interventions have been broadly identified as effective in improving student test scores. First, providing financial or employment incentives to teachers, such as teacher performance pay or contract teacher programs, has been shown to be effective in increasing teacher effort and improving student test scores (Duflo, Hanna, and Ryan 2012; Muralidharan and Sundararaman 2011; Duflo, Dupas, and Kremer 2015; Mbiti et al. 2019). However, countries often struggle to implement these incentive systems nationwide due to bureaucratic or political constraints (Bold et al. 2017, 2018).

The second successful approach is scripted instruction, which provides teachers with detailed lesson plans to address low teacher skills and knowledge. While this approach may be effective in improving very basic skills (Piper and Korda 2011; Piper et al., 2018; Graham and Kelly 2019; Cilliers et al. 2020),[2] concerns have been raised in the United States, where scripted instruction is widespread, that it does not meet the needs of individual students, cannot address complex and higher-level learning, and undermines teacher competence and professionalization (Dresser 2012; Milner 2013).

The third promising approach is teacher coaching, in which experienced coaches visit schools, observe classrooms, provide feedback to teachers, and demonstrate effective teaching techniques. Several experimental studies show that coaching improved teachers' skills and student achievement (Albornoz et al. 2020; Cilliers et al. 2020). However, some countries have struggled to implement coaching effectively nationwide because it is costly and requires a substantial poop of well-trained coaches. Therefore, these interventions alone, while offering partial solutions, may not be sufficient to address the fundamental challenges faced by low- and middle-income countries.

[2] Scripted instruction is a core component of Early Grade Reading (EGR) interventions, a popular literacy program in low- and middle-income countries. However, in their review of 15 EGR interventions, Graham and Kelly (2019) concluded that many of the interventions fail to bring most students to an acceptable reading level, although they can improve simple reading skills in a short period of time.

Compared to these programs, traditional teacher professional development (PD) programs, which have been widely implemented in low- and middle-income countries, have rarely been rigorously evaluated, and many appear ineffective in improving teachers' skills and student learning outcomes in these regions. For instance, Loyalka et al. (2019) demonstrate that a large-scale teacher PD program in China had no impact on student test scores, teacher knowledge, or teaching practices. Popova et al. (2022), comparing large-scale government-funded PD programs with successful PD programs, found that the former differ sharply from the latter in their characteristics. They suggest that successful programs often incur higher per-pupil costs and may not be feasible to scale up nationwide without compromising essential elements. This argument is reinforced by an experimental study in Uganda, where Kerwin and Thornton (2021) show that— while a combined program that includes material inputs, high-quality teacher training, and intensive coaching improved children's reading and writing scores—the impact disappeared when the program removed expensive materials, introduced a cascade training model, and reduced support visits to teachers. Thus, it remains unclear how low- and middle-income countries can improve teachers' skills and knowledge within government systems.

This study also contributes to the discussion on the effectiveness of activity-based instruction in low- and middle-income countries. While activity-based instruction is considered effective for student learning in developed countries, recent studies suggest that this practice has had little or even a negative impact on student test scores in middle-income countries. Berlinski and Busso (2017) found that teacher training to introduce active learning in mathematics may have had a negative impact on student test scores in Costa Rica. De Barros et al. (2023) demonstrate that activity-based instruction had a minimal impact on student test scores, and even the addition of a community monitoring intervention to the instruction yielded no effect in India.

The remaining sections of this paper are organized as follows: Section 2 describes the program and data collection process for the evaluation. Section 3 outlines the evaluation methodology, and Section 4 presents the program's impacts on teaching practice, teacher content knowledge, continuous professional development for teachers, and student learning outcomes. Section 5 discusses the key findings and implications of the paper.

2. Context, Program Description, and Experimental Design

Since 2000, Ethiopia has made significant improvements in access to primary education. Primary school enrollment increased from 5.8 million in 2000 to 16.4 million in 2015, and the net enrollment rate rose from 38.45% to 79.78% during the same period. To accommodate this expansion, the country nearly tripled the number of teachers between 2000 and 2015. However, the quality of education, particularly in mathematics and science, remains a significant challenge. According to the 2011 National Learning Assessment in Ethiopia, nearly half or more of grade 7

and 8 students performed below basic level in mathematics and science (JICA 2012).
To improve the quality of mathematics and science teachers, the Federal Ministry of Education in Ethiopia launched the National Pilot Project for Strengthening Mathematics and Science in 2011, targeting grade seven and eight primary school teachers in mathematics and science. This project is part of a teacher PD initiative called "Strengthening of Mathematics and Science Education (SMASE) in Africa," which aims to shift mathematics and science lessons from teacher-centered to student-centered approaches, introducing activity-based instruction, experiments, and improvisation. The initiative began in Kenya with technical assistance from JICA, and the SMASE project has since expanded to 14 other African countries (Matachi and Kosaka 2017; Ishihara and Kawaguchi 2022).[3]

Based on the SMASE project, we designed two interventions for mathematics and physics teachers and their students.[4] The first intervention was a teacher PD program using the SMASE approach. The training was designed to help teachers understand the concept of active learning and how to implement it in the classroom, focusing on lesson plan development, lesson evaluation, and hands-on activities for mathematics and laboratory techniques in physics. Lesson demonstrations were conducted at a school near the training center. Teachers also learned how to continue PD at their own schools through peer learning activities known as lesson study, in which one teacher conducts a lesson in a real classroom with peer teachers present, followed by a discussion among the teacher and observers about how to improve the lesson.

The second intervention comprised newly developed student workbooks created for this experimental intervention. The SMASE PD program aims to improve student learning in the long term, unlike many initiatives focused on short-term test-score gains. To address the latter need, the workbooks were developed as supplementary resources for countries seeking more immediate improvements in learning outcomes. We hypothesize that even if teaching practices improve, test scores may not improve unless students have the basic skills and knowledge needed to understand lessons and they retain the material taught in class. Based on this hypothesis, the workbooks included (i) basic exercises to review knowledge and skills learned in previous years and (ii) basic and applied exercises to consolidate and apply what was taught in class. Students used the workbooks in supplemental classes (90 minutes per week for mathematics and 60 minutes per week for physics), and teachers monitored each student's progress and explained the answers to the exercises.

[3] Although JICA has ended technical cooperation in SMASE, African countries have continued their own efforts to implement teacher PD programs and have developed a network called SMASE-Africa to promote classroom practices through research, while fostering policy development, networking, collaboration, advocacy and capacity building for education in Africa. As of 2025, 27 countries have joined this network (SMASE-Africa 2025).

[4] In Ethiopia, students in grades 7 and 8 study biology, chemistry, and physics separately, and physics was the lowest-performing subject among the three in the 2011 National Learning Assessment (JICA 2012).

In both interventions, the program covered only the minimum necessary costs for training participants—such as meals, transportation, and accommodation—but did not provide financial incentives. This is a basic principle of SMASE projects in Africa, as it ensures the program's long-term sustainability and nationwide scalability while preventing teachers from becoming dependent on incentives that could hinder their continuous professional development after the interventions. Instead of providing financial incentives, the SMASE programs aim to cultivate intrinsic motivation in teachers, such as the joy of teaching and professional growth (Ishihara 2012; Ishihara and Kawaguchi 2022).

To evaluate these interventions, our experiment planned to cover 335 schools in the Oromia Region that had not previously benefited from the pilot SMASE project. Among these schools, 220 had at least one mathematics teacher and one physics teacher for grade 7 in 2015, while 115 schools had only one teacher for both subjects. In the former case, all schools participated in the intervention in both subjects. In the latter case (referred to henceforth as one-teacher schools), based on the teachers' academic backgrounds, 58 schools took part in the mathematics intervention, and 57 schools in the physics intervention. Before random assignment, 10 strata were established based on administrative zones (five) and whether a school had one teacher or not. Within each stratum, schools were randomly assigned to (i) the Regular PD group (93 schools for mathematics and 92 schools for physics) that received in-service training for mathematics and physics teachers; (ii) the PD-plus-Workbook group (91 schools for mathematics and 92 schools for physics) that received student workbooks in addition to in-service training; or (iii) the control group (92 schools for mathematics and 91 schools for physics).

For both the Regular PD and PD-plus-Workbook groups, PD trainings for teachers in mathematics and physics were conducted at the training center in the Oromia region. In each subject, the training consisted of three cycles, with each cycle comprising five-day sessions (15 days in total), held in February, March, and June 2015.

For the PD-plus-Workbook group, the workbooks were delivered in September 2015, at the beginning of the school year. To ensure that the workbooks were used effectively, three workshops were held: the first workshop took place over three days in June 2015 for administrators and school principals to understand the concept and develop an implementation plan; the second workshop occurred over two days in September 2015 for school principals and teachers to understand the contents of the workbooks and finalize the implementation plan; and the third workshop was held for two days in January 2016 for principals and teachers to share best practices and discuss challenges in using the workbooks.

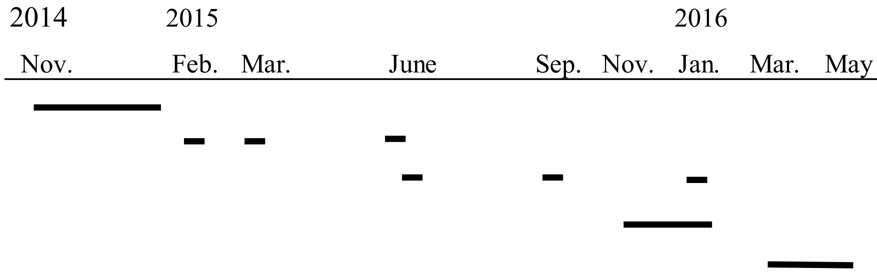
For the evaluation, we collected data through classroom lesson observations, teacher tests (administered only at midline and endline surveys), student tests, and interviews with school principals, teachers, and students. Lesson observations were conducted to analyze how teachers prepared and taught in their classrooms. Trained evaluators assessed classroom instruction using a lesson observation checklist containing 14 criteria, scoring each criterion as “4 = very good,” “3 = good,” “2 = poor,” or “1 = very poor.”[5] Scores were standardized into z-scores by subtracting the control group mean and dividing by the standard deviation for each criterion. The student tests were designed to align with students’ grade levels according to the national curricula and textbooks. The tests for teachers were conducted to evaluate their content knowledge, including questions that were more challenging than those on student tests. Figure 1 illustrates the timeline of the surveys and interventions described above. We conducted the baseline survey from November 2014 to January 2015, the midline survey from November 2015 to February 2016, and the endline surveys from late March to the end of April.[6]

---

[5] The observation checklist is attached in Appendix Figure A1.

[6] The midline survey was scheduled for completion by December 2015, but due to a demonstration against the government that started in the same month, the survey was delayed, and we were unable to survey 19 of the schools.

Figure 1：Timeline of the Surveys and Interventions.



	Control	Regular PD	PD-plus-Workbook
Baseline survey	✓	✓	✓
In-service training		✓	✓
Workshop to utilize workbook			✓
Midline survey	✓	✓	✓
Endline survey	✓	✓	✓

3. Baseline balance and attrition

3.1 Descriptive Statistics and Balance

Table 1 presents summary statistics and balance tests for the three groups, examining school, teacher, and student characteristics from the baseline survey. To evaluate balance, each variable was regressed on treatment assignments, with cluster-standard errors applied at the school level for student characteristics.

Panel A shows the baseline balance of school, teacher, and student characteristics for mathematics teachers, while Panel B presents the baseline balance for physics teachers. School characteristics include the number of students, the number of seventh-grade students, the experience of school principals (measured in years), and the school location (whether it was in a rural or urban area). Teacher characteristics include gender, age, teaching experience, education, teaching practices measured by lesson observations, and school-based professional development (PD) practices (the number of times teachers conducted peer learning activities).[7] Student characteristics include gender, age, and test scores. For mathematics teachers, these characteristics are balanced across the three groups, except for the teacher and the principal experience. The average teacher experience of the PD-plus-Workbook group is nearly 1.2 years higher than the control mean, while the average principal experience in the PD-plus-Workbook group is nearly 0.9 years lower than the Regular PD group average. For physics teachers, all characteristics are balanced across the three groups.

3.2 Attrition

Panel C of Table 1 shows teacher attrition rates categorized by survey round and treatment status. A teacher was considered to have attrited if they participated in the baseline survey as a mathematics or physics grade 7 teacher but did not participate in the endline survey as a grade 8 teacher of the same subject. High teacher attrition was anticipated as primary schools in Ethiopia have eight grades, and many teachers who taught grade 7 during the baseline were not assigned to grade 8 at the midline and endline when the school year changed. For mathematics teachers, the attrition rates for the control group, Regular PD, and PD-plus-Workbook teachers at the endline were 50.0%, 39.8%, and 30.8%, respectively. For physics teachers, the teacher attrition rates for the control, Regular PD, and PD-plus-Workbook groups were 68.1%, 47.8%, and 30.4%, respectively.

[7] Many schools in Ethiopia, whether in the treatment or control group, conduct school-based PD practices. Teachers in treatment groups were encouraged to continue PD activities at the school level after the SMASE program, but these school-level PD practices were not directly implemented by our experiment.

Table 1: Pretreatment Balance and Attrition

	Control	PD	Workbook (WB)	PD vs Control		WB vs Control		WB vs PD	
	Mean	Mean	Mean	Mean Difference	SE	Mean Difference	SE	Mean Difference	SE
Panel A: Mathematics Teacher									
School and teacher characteristics									
School is located in rural area	0.88	0.839	0.868	-0.041	(0.051)	-0.012	(0.049)	0.029	(0.052)
Number of students	652.424	619.548	629.143	-32.876	(48.626)	-23.281	(57.975)	9.595	(54.851)
School director experience (years)	4.37	4.774	3.901	0.404	(0.508)	-0.469	(0.446)	-0.873*	(0.481)
Female Teacher (Yes/no)	0.076	0.097	0.088	0.021	(0.042)	0.012	(0.041)	-0.009	(0.043)
Teacher age	26.207	26.419	26.989	0.212	(0.527)	0.782	(0.655)	0.57	(0.689)
Teacher experience (years)	4.891	5.742	6.066	0.851	(0.577)	1.175*	(0.661)	0.324	(0.701)
Teacher has diploma or degree	0.967	0.989	0.978	0.022	(0.022)	0.011	(0.024)	-0.011	(0.045)
Teacher majored in mathematics (Yes/no)	0.891	0.903	0.868	0.012	(0.045)	-0.023	(0.048)	-0.035	(0.047)
Teaching practice (index)	0	-0.006	0.044	-0.006	(0.076)	0.044	(0.078)	0.05	(0.069)
School-based PD	1.391	1.228	1.341	-0.163	(0.190)	-0.05	(0.191)	0.113	(0.199)
Observations	92	93	91						
p-value of joint F-test				0.417		0.246		0.444	
Student characteristics									
Female student	0.465	0.486	0.491	0.021	(0.025)	0.026	(0.023)	0.005	(0.026)
Age	13.998	13.951	13.94	-0.047	(0.110)	-0.058	(0.099)	-0.011	(0.108)
Mathematics Test Score	0	-0.005	0.059	-0.005	(0.063)	0.059	(0.067)	0.064	(0.067)

Observations	918	917	903						
p-value of joint F-test				0.852		0.449		0.824	
Panel B: Physics Teacher									
School and teacher characteristics									
School is located in rural area	0.879	0.837	0.88	-0.042	(0.052)	0.001	(0.048)	0.043	(0.052)
Number of students	647.89	605.859	601.696	-42.031	(48.324)	-46.194	(59.068)	-4.163	(55.668)
School director experience (years)	4.615	4.576	4.098	-0.039	(0.511)	-0.517	(0.464)	-0.478	(0.478)
Female Teacher (Yes/no)	0.165	0.109	0.109	-0.056	(0.051)	-0.056	(0.051)	0	(0.046)
Teacher age	26.473	25.598	26.054	-0.875	(0.661)	-0.419	(0.664)	0.456	(0.580)
Teacher experience (years)	5.484	4.5	5.196	-0.984	(0.614)	-0.288	(0.627)	0.696	(0.600)
Teacher has diploma or degree	0.978	0.967	0.967	-0.011	(0.024)	-0.011	(0.024)	0	(0.026)
Teacher majored in Physics (Yes/no)	0.341	0.293	0.38	-0.048	(0.069)	0.039	(0.071)	0.087	(0.070)
Teaching practice (index)	0	-0.102	-0.035	-0.102	(0.078)	-0.035	(0.078)	0.067	(0.073)
School-based PD	1.478	1.185	1.359	-0.293	(0.191)	-0.119	(0.195)	0.174	(0.191)
Observations	91	92	92						
p-value of joint F-test				0.130		0.951		0.390	
Student characteristics									
Female student	0.458	0.476	0.485	0.018	(0.024)	0.027	(0.026)	0.009	(0.028)
Age	13.954	13.957	13.92	0.003	(0.108)	-0.034	(0.091)	-0.037	(0.108)
Physics Test Score	0	-0.139	-0.061	-0.139**	(0.058)	-0.061	(0.057)	0.078	(0.061)
Observations	904	906	903						
p-value of joint F-test				0.127		0.534		0.589	

Panel C: Attrition Rate									
Mathematics Teacher	0.500	0.398	0.308	-0.102	(0.073)	-0.192***	(0.072)	-0.090	(0.071)
Observations	92	93	91						
Physics Teacher	0.681	0.478	0.304	-0.203***	(0.072)	-0.377***	(0.069)	-0.174**	(0.071)
Observations	91	92	92						

Notes: Panels A and B show the pretreatment balance of school, teacher, and student characteristics in mathematics and physics, respectively. For student characteristics, standard errors are clustered at the school level. The number of observations for teaching practice in the control, PD, and Workbook groups in mathematics is 91, 91, and 90, respectively, and that in physics is 89, 92, and 91, respectively (the numbers are lower than those of the other teacher characteristics because these teachers did not have class on the survey date). Test scores are normalized so that the mean and standard deviation of the control group are zero and one. Panel C shows the attrition rate of teachers.
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

4. Empirical strategy

The effects of the two interventions can be estimated using the following equation:

$$y_{ij} = \alpha + \beta_1 PD_j + \beta_2 WORKBOOK_j + X_{ij}\gamma + \varepsilon_{ij}$$

where y_j or y_{ij} is an outcome of a teacher j or a student i of teacher j; PD_j is a dummy equal to 1 if a teacher j is assigned into the Regular PD group and receives teacher training but does not have the workbook intervention, and zero otherwise; $WORKBOOK_j$ is a dummy equal to one if a teacher j is assigned into the PD-plus-Workbook group and receives the workbook intervention as well as teacher training, and zero otherwise; and, X_{ij} is a vector of control variables, including strata dummies, school location (rural or urban area), number of students, principal's experience, teacher gender, teacher age, teacher experience, teacher education, teacher major, student gender, student age, baseline outcome value when available, and whether the school was affected by the demonstration against the government that began in December 2015.[8] The parameters of interest are β_1, which is the average treatment effect of the Regular PD, and β_2, which is the average treatment effect of the PD-plus-Workbook. Standard errors are clustered at the school level for student test scores.

Given the high attrition rate of teachers, the treatment effects were estimated by using inverse probability weighting with the same equation above. To estimate the probability of response, strata dummies were used along with school and teacher characteristics.

5. Results

5.1 Teaching Quality

Table 2 presents the impacts on teaching practice at the midline (six months after the PD training and three months after the workbook introduction) and at the endline (ten months after the PD training and seven months after the workbook introduction).

Columns 1–2 and 3–4 present the results of the main specification from lesson observations at the midline and endline, respectively. The findings indicate that, at the midline, both the Regular PD group and the PD-plus-Workbook group showed improvements in teaching practices overall in both subjects. For the mean index, the increases for the Regular PD and the PD-plus-Workbook were 0.344 and 0.298 standard deviations in mathematics, and 0.473 and 0.432 in physics, respectively. By the endline, the impacts grew larger. The increase in the mean index for the Regular PD and the PD-plus-Workbook was 0.582 and 0.637 standard deviations in mathematics, and 0.494 and 0.771 standard deviations in physics, respectively. When comparing Regular PD and PD-plus-Workbook teachers, there was little difference in mathematics; however, PD-plus-

[8] Regarding baseline outcome data, teaching quality, continuous professional development, and student test scores are available, but teacher knowledge is unavailable.

Workbook teachers outperformed their Regular PD counterparts by 0.277 standard deviations in physics at the endline. In addition, columns 5–8 present the results of estimation using inverse probability weighting. These results indicate that the weighted estimates closely align with the primary specification estimates.

Table 2: Treatment Effects on Teaching Practice

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Main Specification				Inverse Probability Weighting			
	Midline		Endline		Midline		Endline	
	Math	Physics	Math	Physics	Math	Physics	Math	Physics
β_1: Regular PD	0.344***	0.473***	0.582***	0.494***	0.332***	0.476***	0.611***	0.460***
	(0.098)	(0.133)	(0.118)	(0.111)	(0.096)	(0.134)	(0.121)	(0.113)
β_2: PD-plus-Workbook	0.298***	0.432***	0.637***	0.771***	0.276***	0.459***	0.593***	0.778***
	(0.095)	(0.121)	(0.117)	(0.128)	(0.097)	(0.131)	(0.113)	(0.138)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	172	151	165	141	172	151	165	141
R-squared	0.210	0.263	0.289	0.387	0.198	0.260	0.283	0.380
p-value (β_2-β_1=0)	0.663	0.728	0.624	0.009	0.594	0.885	0.875	0.004

Notes: Regressions control for strata dummies, school location (rural or urban area), number of students, principal's experience, teacher gender, teacher age, teacher experience, teacher education, teacher major, and baseline outcome value when available. Robust standard errors are in parentheses.

*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

5.2 Teacher Content Knowledge

Table 3 shows the impacts on teachers' content knowledge at both the midline and endline. The results indicate positive effects on teacher knowledge for both treatment groups at the midline, although a statistically significant impact is observed only for PD-plus-Workbook teachers in mathematics. Similar to teaching practice, the effect was stronger at the endline. The Regular PD enhanced teacher knowledge by 0.383 standard deviations in mathematics and 0.381 in physics, while PD-plus-Workbook improved knowledge by 0.479 standard deviations in mathematics and 0.626 standard deviations in physics. As with teaching quality, the impact on PD-plus-Workbook teachers was more pronounced than that on Regular PD teachers in physics. The weighted estimates, shown in columns 5–8, also yield similar results.

Table 3: Treatment Effects on Teacher Content Knowledge

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Main Specification				Inverse Probability Weighting			
	Midline		Endline		Midline		Endline	
	Math	Physics	Math	Physics	Math	Physics	Math	Physics
β_1: Regular PD	0.009	0.411	0.383**	0.381*	-0.015	0.189	0.363*	0.410*
	(0.209)	(0.258)	(0.201)	(0.228)	(0.217)	(0.271)	(0.191)	(0.214)
β_2: PD-plus-Workbook	0.544***	0.403	0.479**	0.626***	0.615***	0.228	0.507***	0.666***
	(0.185)	(0.267)	(0.181)	(0.215)	(0.193)	(0.284)	(0.177)	(0.198)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	167	148	165	141	167	148	165	141
R-squared	0.164	0.161	0.162	0.194	0.221	0.149	0.206	0.213
p-value (β_2-β_1=0)	0.009	0.974	0.559	0.157	0.003	0.871	0.363	0.125

Notes: Regressions control for strata dummies, school location (rural or urban area), number of students, principal's experience, teacher gender, teacher age, teacher experience, teacher education, and teacher major. Robust standard errors are in parentheses.
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

5.3 Continuous Professional Development

Table 4 shows the impact of teachers' participation in continuous professional development practices at their schools to enhance teaching quality. SMASE training encouraged teachers to implement these practices at their own schools, enabling them to continue learning and improving their teaching skills and knowledge.

At the baseline and midline, surveyors asked teachers how often they had participated in collaborative learning at their schools during the previous school year. At the endline, surveyors inquired about their participation from the beginning of the current school year up to the survey date.[9] The results indicate that both treatment groups increased this activity at the midline and endline. Columns 1 and 2 show that, at the midline, the Regular PD mathematics and physics teachers increased their activities by 1.216 times (127%) and 0.804 times (94%), respectively. Meanwhile, the PD-plus-Workbook mathematics and physics teachers increased their collaborative learning by 1.212 times (122%) and 1.285 times (150%), respectively, relative to the control means of 0.957 times in mathematics and 0.857 times in physics.

Columns 3 and 4 show that, at the endline, the Regular PD mathematics and physics teachers increased their activities by 0.412 times (44%) and 0.126 times (13%), respectively, and the PD-plus-Workbook mathematics and physics teachers increased theirs by 0.433 times (46%) and 0.469 times (47%), respectively, relative to the control mean of 0.936 times in mathematics and 1.000 times in physics. The impacts appear to be smaller at the endline than at the midline, possibly due to the shorter period covered at the endline.

[9] As a result, the covered period was nearly three months shorter at the endline than at the baseline and midline.

Table 4: Treatment Effects on Continuous Teacher PD Practice

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Main Specification				Inverse Probability Weighting			
	Midline		Endline		Midline		Endline	
	Math	Physics	Math	Physics	Math	Physics	Math	Physics
β_1: Regular PD	1.216***	0.804***	0.412*	0.126	1.232***	0.846***	0.366	0.201
	(0.229)	(0.276)	(0.241)	(0.283)	(0.230)	(0.292)	(0.242)	(0.270)
β_2: PD-plus-Workbook	1.212***	1.285***	0.433*	0.469*	1.251***	1.294***	0.384*	0.549**
	(0.222)	(0.264)	(0.224)	(0.273)	(0.225)	(0.279)	(0.219)	(0.260)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	172	151	165	141	172	151	165	141
R-squared	0.318	0.302	0.149	0.241	0.334	0.283	0.132	0.263
Control Mean	0.957	0.857	0.936	1.000	0.940	0.891	0.982	0.965
p-value (β_2-β_1=0)	0.985	0.048	0.924	0.119	0.948	0.074	0.862	0.103

Notes: Regressions control for strata dummies, school location (rural or urban area), number of students, principal's experience, teacher gender, teacher age, teacher experience, teacher education, teacher major, and baseline outcome value when available. Robust standard errors are in parentheses.
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

5.4 Student test scores

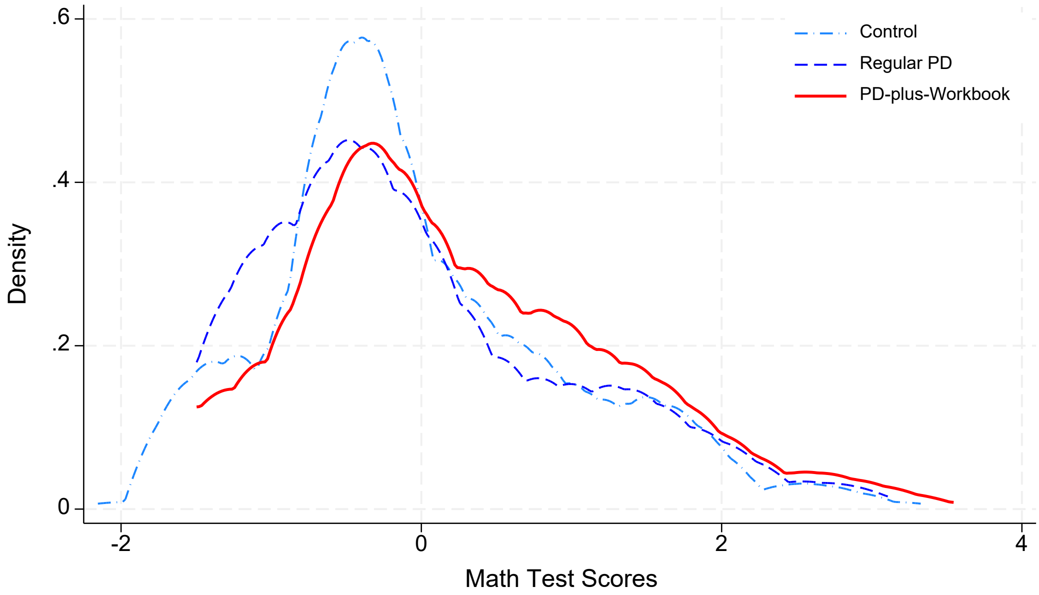
Table 5 presents the estimated impacts on student test scores, and Figure 2 illustrates the distribution of endline test scores across the three groups. The results indicate that, while the regular professional development (PD) alone did not improve student test scores, the PD-plus-Workbook approach yielded a statistically significant impact on physics. In the main specification, students in the PD-plus-Workbook group had an average test score that was 0.251 standard deviations higher in physics compared to the control group students. The average score of the PD-plus-Workbook group students was also higher in mathematics by 0.143 standard deviations, although this difference was not statistically significant. Estimates using inverse probability weighting show similar results, except for the average mathematics score of the Regular PD group at the midline, which is significantly lower than that of the control group.

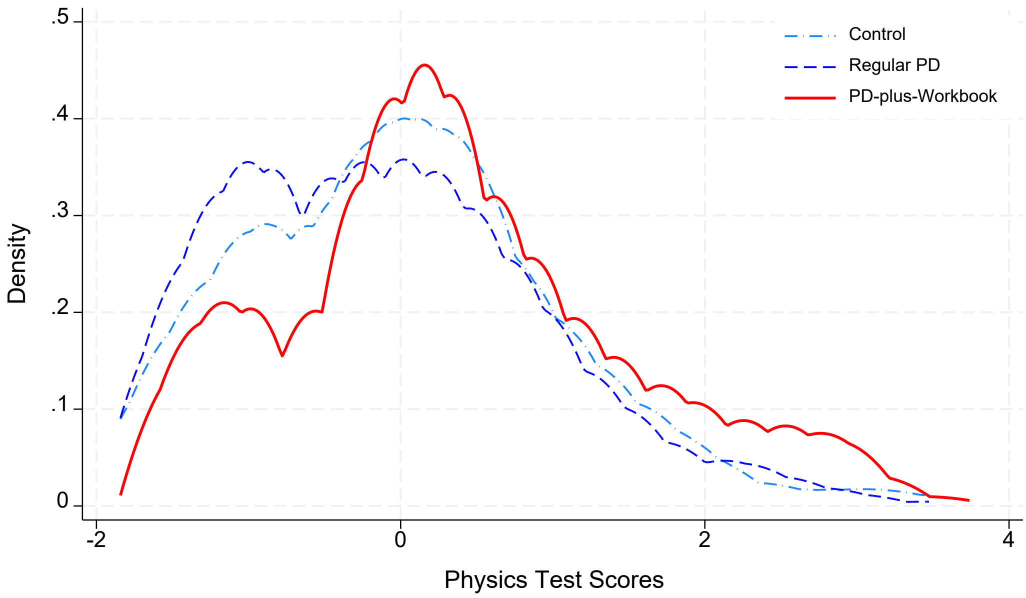
Table 5: Treatment Effects on Student Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Main Specification				Inverse Probability Weighting			
	Midline		Endline		Midline		Endline	
	Math	Physics	Math	Physics	Math	Physics	Math	Physics
β_1: Regular PD	-0.112	-0.032	0.004	-0.064	-0.146*	-0.077	-0.023	0.009
	(0.092)	(0.125)	(0.102)	(0.127)	(0.083)	(0.119)	(0.099)	(0.131)
β_2: PD-plus-Workbook	0.010	0.112	0.143	0.250**	0.013	0.116	0.157	0.334**
	(0.089)	(0.101)	(0.103)	(0.124)	(0.084)	(0.102)	(0.099)	(0.140)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,687	1,480	1,589	1,356	1,687	1,480	1,589	1,356
R-squared	0.134	0.134	0.116	0.119	0.140	0.140	0.123	0.123
p-value (β_2-β_1=0)	0.128	0.198	0.125	0.003	0.039	0.068	0.046	0.002

Notes: Regressions control for strata dummies, school location (rural or urban area), number of students, principal's experience, teacher gender, teacher age, teacher experience, teacher education, teacher major, student gender, student age, and baseline test scores. Test scores are normalized so that the mean and standard deviation of the control group are zero and one. Standard errors are clustered at the school level.
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

Figure 2: Distribution of Student Test Scores at the Endline





5.5 Heterogeneous Effects by Teacher Experience

Table 6 presents the heterogeneous impact on teachers by teaching experience, defined as whether a teacher had more than five years of experience at baseline, which divided the sample roughly equally. The results indicate that both interventions were more effective for less experienced teachers in both subjects. Regular professional development (PD) had a positive influence on teaching practices and knowledge, while PD-plus-Workbook had a positive impact on all outcomes for less experienced teachers. In contrast, for more experienced teachers, both interventions only positively influenced teaching practices.

5.6 Heterogeneous Effects by Student Baseline Test Scores

Table 7 shows the heterogeneous impacts on student test scores by categorizing students into quintiles based on their average baseline test scores. In the Regular PD group, there was no statistically significant difference across any quintile. However, for the PD-plus-Workbook group, the effects on higher-performing students were both positive and statistically significant, with a 0.452 standard deviation for the second-highest quintile in mathematics, a 0.382 standard deviation for the second-highest quintile in physics, and a 0.801 standard deviation for the highest quintile in physics.

Table 6: Heterogeneous Treatment Effects by Teacher Experience

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Five year or less experience						More than five year experience					
	Teaching Practice		Content Knowledge		Student Test Scores		Teaching Practice		Content Knowledge		Student Test Scores	
	Math	Physics	Math	Physics	Math	Physics	Math	Physics	Math	Physics	Math	Physics
β_1: Regular PD	0.734***	0.735***	0.495	0.834**	-0.117	0.100	0.503***	0.339**	0.328	0.011	0.024	-0.122
	(0.211)	(0.188)	(0.297)	(0.340)	(0.116)	(0.171)	(0.142)	(0.154)	(0.290)	(0.385)	(0.154)	(0.184)
β_2: PD-plus-Workbook	0.885***	1.028***	0.678**	0.896**	0.228*	0.395**	0.462**	0.730***	0.382	0.427	0.020	0.127
	(0.184)	(0.178)	(0.292)	(0.347)	(0.129)	(0.190)	(0.188)	(0.167)	(0.230)	(0.319)	(0.144)	(0.143)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	78	75	78	75	745	710	87	66	87	66	844	646
R-squared	0.421	0.566	0.377	0.317	0.150	0.108	0.365	0.458	0.254	0.279	0.082	0.100
Control Mean	-0.030	-0.294	-0.019	-0.165	0.007	-0.156	0.029	0.211	0.022	0.116	-0.086	0.101
p-value (β_2-β_1=0)	0.001	0.000	0.100	0.017	0.314	0.560	0.001	0.033	0.262	0.978	0.874	0.512

Notes: Regressions control for strata dummies, school location (rural or urban area), number of students, principal's experience, teacher gender, teacher age, teacher experience, teacher education, teacher major, student gender, student age, and baseline outcome value when available. Regressions use robust standard errors for teaching practice and teacher knowledge, and clustered standard errors at the school level for student test scores.

*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

Table 7: Heterogeneous Effects on Student Test Scores

	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
			Math					Physics		
	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
β_1: Regular PD	0.017	-0.108	0.304	-0.059	-0.141	-0.040	-0.297	0.031	-0.078	0.436
	(0.165)	(0.183)	(0.257)	(0.264)	(0.287)	(0.279)	(0.308)	(0.358)	(0.192)	(0.394)
β_2: PD-plus-Workbook	0.084	0.205	0.256	0.452**	-0.110	0.254	0.254	0.203	0.382*	0.801*
	(0.177)	(0.164)	(0.287)	(0.217)	(0.309)	(0.270)	(0.290)	(0.292)	(0.193)	(0.404)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	189	219	86	158	141	206	119	98	170	122
R-squared	0.188	0.211	0.261	0.239	0.225	0.233	0.249	0.215	0.200	0.269
Control mean	-0.226	-0.135	-0.13	0.129	0.835	-0.228	-0.107	0.018	-0.074	0.194
p-value (β_2-β_1=0)	0.748	0.077	0.969	0.028	0.933	0.090	0.039	0.650	0.025	0.203
p-value (equality of β_2)			0.316					0.668		

Notes: Students are divided by quintile based on their baseline test scores, from lowest (Quintile 1) through highest (Quintile 5) in each subject. Regressions control for strata dummies, school location (rural or urban area), number of students, principal's experience, teacher gender, teacher age, teacher experience, teacher education, teacher major, student gender, student age, and baseline test scores when available. Robust standard errors are clustered at the school level and are in parentheses. 'p-value (equality of β_2)' shows the p-value of a Chow test of equality of β_2 across quintiles.
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

5.7 Robustness to Attrition

A limitation of this study is the high teacher attrition rate, which could lead to biased estimates of the treatment effects. To examine whether teacher attrition is selective, the methodology proposed by Ghanem et al. (2024) was applied, with results shown in Table 8. "IVal-R," or internal validity for the respondent subpopulation, tests the equality of the main baseline outcomes among treatment and control teachers who participated in the endline survey (respondents) as well as across treatment and control teachers who attritted (attritors). The tests do not reject the null hypothesis of equality of the baseline outcomes, except for physics student test scores, conditional on response status. This suggests that the main results in the previous sections have internal validity for the respondent subpopulation. "IVal-P," or internal validity for the study population, tests the equality of the main baseline outcomes among all six treatment/response combinations. These tests cannot reject the null hypothesis of equality, indicating that the results also have internal validity for the study population.

Table 8: Selective Attrition Test

		Mean Baseline Outcome by group						Test of IVal-R	Test of IVal-P
		Respondents			Attritors			*p*-value	
	Obs.	Control	Regular PD	PD-plus-WB	Control	Regular PD	PD-plus-WB		
Mathematics teacher									
Teaching practice (index)	272	-0.023	0.007	0.032	0.025	-0.039	-0.026	0.754	0.902
School-based PD	275	1.522	1.200	1.317	1.261	1.548	1.270	0.759	0.833
Student test scores	2738	0.037	0.041	0.075	-0.037	-0.074	0.022	0.914	0.355
Physics teacher			.						
Teaching practice (index)	272	0.057	-0.056	-0.037	-0.151	0.073	-0.030	0.780	0.758
School-based PD	274	1.321	1.125	1.562	1.250	1.393	0.893	0.111	0.124
Student test scores	2713	0.007	-0.197	-0.047	-0.003	-0.075	-0.092	0.068	0.146

Notes: "IVal-R" tests the equality of the main baseline outcomes among treatment and control teachers who participated in the endline survey (respondents) as well as between the treatment and control teachers who attritted (attritors). "IVal-P" tests the equality of the main baseline outcomes across all six treatment/response combinations. These tests are implemented based on the methodology proposed by Ghanem et al. (2024). Regressions use robust standard errors for teaching practice and PD, and clustered standard errors at the school level for student test scores. The tests were conducted using the Stata command *attregtest*.

Table 9 also presents robustness to attrition by re-estimating the treatment effects in alternative ways. Columns 1 and 2 repeat the main results from Tables 2, 3, and 5.

Columns 3 and 4 calculate the lower bounds by imputing outcome values for the control attritors using the group means plus 0.1 and 0.2 standard deviations, respectively, and for the treatment attritors using the group means minus 0.1 and 0.2 standard deviations, respectively. Columns 6 and 7 calculate the upper bounds by imputing the outcomes for the control group attritors with the group means minus 0.1 and 0.2 standard deviations and imputing the outcomes of the treatment group attritors with the group means plus 0.1 and 0.2 standard deviations. The results indicate that both treatment groups have significant impacts on teaching practice and teacher knowledge, even at the lower bounds, except for Regular PD's impacts on teacher knowledge under extreme scenarios (±0.2 standard deviations). For student test scores, however, the impacts become negative.

Columns 5 and 8 present Lee bounds (Lee 2009), which represent the smallest and largest possible treatment effects by trimming the upper and lower tails of the outcome distribution based on the number of individuals induced to be selected by the treatment. While the lower bounds of teacher pedagogy remain statistically significant across both subjects, the lower bounds for the other outcomes become insignificant. In summary, the results of teaching practice are robust when controlling for attrition, even under extreme scenarios, whereas the effects of other outcomes disappear under such situations.

Table 9: Alternative Estimation of Treatment Effects to Deal with Attrition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Main specification	Inverse Probability Weighting	Lower bound			Upper bound		
			±.1 SD	±.2 SD	Lee	±.1 SD	±.2 SD	Lee
Mathematics								
Teaching Practice								
Regular PD	0.582***	0.611***	0.448***	0.356***	0.320**	0.631***	0.723***	0.704***
	(0.118)	(0.121)	(0.061)	(0.063)	(0.138)	(0.060)	(0.061)	(0.147)
PD-plus-Workbook	0.637***	0.593***	0.536***	0.454***	0.375***	0.699***	0.781***	0.911***
	(0.117)	(0.113)	(0.071)	(0.073)	(0.134)	(0.072)	(0.073)	(0.132)
Teacher Knowledge								
Regular PD	0.383**	0.363*	0.195*	0.103	0.060	0.379***	0.471***	0.536**
	(0.191)	(0.191)	(0.104)	(0.105)	(0.258)	(0.104)	(0.105)	(0.248)
PD-plus-Workbook	0.479**	0.507***	0.337***	0.255**	0.246	0.501***	0.582***	0.973***
	(0.185)	(0.177)	(0.104)	(0.105)	(0.220)	(0.104)	(0.104)	(0.184)
Student Test Scores								
Regular PD	0.004	-0.033	-0.152***	-0.248***	-0.382***	0.038	0.133**	0.174**
	(0.101)	(0.100)	(0.054)	(0.055)	(0.092)	(0.053)	(0.054)	(0.084)
PD-plus-Workbook	0.136	0.156	-0.003	-0.087	-0.414***	0.165***	0.249***	0.508***
	(0.105)	(0.100)	(0.061)	(0.062)	(0.076)	(0.060)	(0.061)	(0.080)
Physics								

Teaching Practice								
Regular PD	0.494***	0.460***	0.376***	0.259***	0.205*	0.611***	0.728***	0.748***
	(0.111)	(0.113)	(0.047)	(0.048)	(0.123)	(0.047)	(0.048)	(0.146)
PD-plus-Workbook	0.771***	0.778***	0.662***	0.562***	0.287**	0.863***	0.963***	1.257***
	(0.128)	(0.138)	(0.058)	(0.059)	(0.134)	(0.058)	(0.060)	(0.154)
Teacher Knowledge								
Regular PD	0.381*	0.410*	0.268***	0.151*	-0.106	0.503***	0.620***	0.926***
	(0.213)	(0.214)	(0.087)	(0.088)	(0.242)	(0.087)	(0.088)	(0.258)
	0.626***	0.666***	0.490***	0.390***	-0.148	0.690***	0.790***	1.300***
PD-plus-Workbook	(0.206)	(0.198)	(0.092)	(0.093)	(0.239)	(0.091)	(0.091)	(0.244)
Student Test Scores								
Regular PD	-0.065	0.004	-0.146***	-0.266***	-0.697***	0.093*	0.213***	0.566***
	(0.127)	(0.131)	(0.051)	(0.053)	(0.096)	(0.051)	(0.053)	(0.105)
PD-plus-Workbook	0.251**	0.332**	0.210***	0.105*	-0.540***	0.420***	0.525***	1.222***
	(0.124)	(0.140)	(0.056)	(0.057)	(0.078)	(0.056)	(0.058)	(0.100)

Notes: Columns 1 and 2 repeat the main results from Tables 2, 3 and 5. Columns 3 and 4 calculate the lower bounds by imputing outcome values for the control attritors with the group means plus 0.1 and 0.2 standard deviations, respectively, and for the treatment attritors with the group means minus 0.1 and 0.2 standard deviations, respectively. Columns 6 and 7 calculate the upper bounds by imputing the outcomes of the control group attritors with the group mean minus 0.1 and 0.2 standard deviations and imputing the outcomes of the treatment group attritors with the group means plus 0.1 and 0.2 standard deviations. Columns 5 and 8 calculate Lee bounds by trimming the upper and lower tails in each treatment group (Lee 2009).
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

6. Discussion and Conclusion

This study evaluated the impacts of a primary school teacher PD program and supplemental student workbooks in mathematics and physics. The study finds that the teacher PD program improved teaching practices and teacher content knowledge but did not enhance student test scores. When student workbooks were provided as a supplement to teacher PD, however, student test scores improved. Although this program did not offer participants financial or promotion-linked incentives during or after the training, teachers continued their efforts to improve their practice, which may have reinforced the impacts of the program.

The impacts of the interventions were stronger for less-experienced teachers, but the effect of the workbooks was weaker for low-performing students. This finding suggests that the content of the interventions aligns with the needs of less experienced teachers, while the level of the workbooks, designed to adhere to the curriculum, was challenging for low-performing students. A similar finding was reported in Kenya, where the provision of textbooks was effective only for high-achieving students (Glewwe, Kremer, and Moulin 2009). To tackle the learning crisis in Africa, where more than half of students are not achieving basic skills (UNESCO Institute for Statistics 2017), the material in teacher training and workbooks, or even the national curriculum and textbooks, may need to be revised so that low-achieving students can better comprehend what they are being taught in class.

The results of this study raise several policy-related questions. First, why did the government-implemented teacher PD program in Ethiopia improve teaching practice, while the government-implemented PD evaluated by Loyalka et al. (2019) in China had no impact? One explanation may lie in differences in program characteristics. Loyalka et al. (2019) attribute a major reason for the lack of impact of professional development (PD) in the Chinese intervention to its content, which was overly theoretical and difficult for teachers to implement, and its delivery, which was rote and passive. In contrast, the PD in Ethiopia emphasized practice over theory: the training taught teachers how to conduct hands-on activities and laboratory techniques, and included extensive discussion and classroom demonstrations.

The second question concerns why improved teaching quality did not translate into better student test scores. Learning theory and empirical evidence suggest that teaching quality does not directly affect student achievement but rather has an indirect effect when students utilize learning opportunities (Praetorius et al. 2018; Christ et al. 2022). In Ethiopia, workbooks may have increased learning opportunities by extending students' learning time and helping them retain material covered in class.

Third, can the workbooks alone improve student test scores in the absence of teacher PD? Although substantial evidence indicates that simply providing inputs such as textbooks and materials may not enhance student learning (Glewwe et al. 2004; Glewwe, Kremer, and Moulin 2009; Sabarwal, Evans, and Marshak 2014; Kozuka 2023), the workbook intervention in Ethiopia not only provided workbooks but also integrated them with teacher assistance in supplementary classes. Thus, if revised for a stand-alone intervention, the workbook intervention could have improved student test scores without teacher PD. However, its impact may be reduced if there are complementarities between the PD and the workbooks, which our program aimed to achieve.[10] We did not create an additional treatment group to test the impact of the single workbook intervention because our goal was not to improve student test scores in the short term at the expense of improving teaching quality.

Finally, should low- and middle-income countries continue implementing teacher PD programs, even if they do not directly improve student test scores? This study alone cannot answer this question; however, if teacher quality is the key to student learning and the learning theory above is correct, then teacher PD should remain a long-term strategy. More fundamentally, the prevailing research trend, which focuses primarily on student test scores, may need to be reconsidered, as teachers should support student growth across multiple dimensions beyond test performance and should be able to respond to student needs rather than simply teaching to the test (Milner 2013).

Low- and middle-income countries face the challenging task of tackling the learning crisis for children today while simultaneously developing their education systems for future generations (World Bank 2018). Future research should provide policymakers with evidence that informs both short-term and long-term solution

[10] For example, Mbiti et al. (2019) show that the test score improvements from a grant program and a teacher performance pay program were significantly larger when implemented together than when implemented separately in Tanzania, suggesting complementarities between the two programs.

References

Albornoz, F., M. V. Anauati, M. Furman, M. Luzuriaga, M. E. Podesta, and I. Taylor. 2020. “Training to Teach Science: Experimental Evidence from Argentina.” World Bank Economic Review, 34 (2): 393–417.

Berlinski, S., and M. Busso. 2017. Challenges in Educational Reform: An Experiment on Active Learning in Mathematics. Economics Letters, 156: 172–175.

Bold, T., D. Filmer, G. Martin, E. Molina, B. Stacy, C. Rockmore, J. Svensson, and W. Wane. 2017. “Enrollment Without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa.” Journal of Economic Perspectives, 31 (4): 185–204.

Bold, T., M. Kimenyi, G. Mwabu, and J. Sandefur. 2018. “Experimental Evidence on Scaling Up Education Reforms in Kenya.” Journal of Public Economics, 168: 1–20.

Chetty, R., J. N. Friedman, and J. E. Rockoff. 2014. “Measuring the Impacts of Teachers II: Teacher Value-added and Student Outcomes in Adulthood.” American Economic Review, 104 (9): 2633–2679.

Christ, A. A., V. Capon-Sieber, U. Grob, and A. K. Praetorius. 2022. “Learning Processes and Their Mediating Role between Teaching Quality and Student Achievement: A Systematic Review.” Studies in Educational Evaluation, 75: 101209.

Cilliers, J., B. Fleisch, C. Prinsloo, and S. Taylor. 2020. “How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching.” Journal of Human Resources, 55 (3): 926–962.

De Barros, A., J. Fajardo-Gonzalez, P. Glewwe, A. Sankar. 2024. “The Limitations of Activity-based Instruction to Improve the Productivity of Schooling.” Economic Journal, 134 (659): 959–984.

Dresser, R. 2012. “The Impact of Scripted Literacy Instruction on Teachers and Students.” Issues in Teacher Education, 21 (1): 71–87.

Duflo, E., P. Dupas, and M. Kremer. 2015. “School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools.” Journal of Public Economics, 123: 92–110.

Duflo, E., R. Hanna, and S. P. Ryan. 2012. “Incentives Work: Getting Teachers to Come to School.” American Economic Review, 102 (4): 1241–1278.

Evans, D. K., and A. Popova. 2016. “What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews.” World Bank Research Observer, 31 (2): 242–270. https://doi.org/10.1093/wbro/lkw004.

Ghanem, D., S. Hirshleifer, and K. Ortiz-Becerra. 2024. “Testing Attrition Bias in Field Experiments.” Journal of Human Resources. https://doi.org/10.3368/jhr.0920-11190R2

Glewwe, P., M. Kremer, and S. Moulin. 2009. “Many Children Left Behind? Textbooks and Test Scores in Kenya.” American Economic Journal: Applied Economics, 1 (1): 112–135. https://doi.org/10.1257/app.1.1.112.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz. 2004. “Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya.” Journal of Development Economics, 74 (1): 251–268. https://doi.org/10.1016/j.jdeveco.2003.12.010.

Graham, J., and S. Kelly. 2019. “How Effective Are Early Grade Reading Interventions? A Review of the Evidence.” Educational Research Review, 27: 155–175.

Hanushek, E. A., and S. G. Rivkin. 2010. “Generalizations about Using Value-added Measures of Teacher Quality.” American Economic Review, 100 (2): 267–271.

Ishihara, S. 2012. “Case 1 Network-type Cooperation: Strengthening of Mathematics and Science Education in Western, Eastern, Central, and Southern Africa (SMASE-WECSA)

Network." In Scaling up South-South and Triangular Cooperation, edited by Hiroshi Kato, 83–106. Tokyo: JICA Research Institute.

Ishihara, S., and J. Kawaguchi. 2022. "Improvement of Teachers' Classroom Practices for Quality Teaching and Learning: The Japanese Approaches." In Japan's International Cooperation in Education. Education in the Asia-Pacific Region: Issues, Concerns and Prospects, vol 63, edited by N. Kayashima, K. Kuroda, and Y. Kitamura. Singapore: Springer. https://doi.org/10.1007/978-981-16-6815-9_5.

Japan International Cooperation Agency (JICA). 2012. "Basic Education Sector Analysis Report, Ethiopia." https://openjicareport.jica.go.jp/pdf/12083135.pdf.

Kerwin, J. T., and R. L. Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." Review of Economics and Statistics, 103 (2): 251–264.

Kozuka, E. 2023. "Enlightening Communities and Parents for Improving Student Learning: Evidence from Niger." Economics of Education Review, 94: 102396.

Kremer, M., E. Miguel, and R. Thornton. 2009. "Incentives to Learn." Review of Economics and Statistics, 91 (3): 437–456. https://doi.org/10.1162/rest.91.3.437.

Lee, D. S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." Review of Economic Studies, 76 (3): 1071–1102. https://doi.org/10.1111/j.1467–937X.2009.00536.x.

Loyalka, P., A. Popova, G. Li, and Z. Shi. 2019. "Does Teacher Training Actually Work? Evidence from a Large-scale Randomized Evaluation of a National Teacher Training Program." American Economic Journal: Applied Economics, 11 (3): 128–154.

Matachi, A., and M. Kosaka. 2017. "JICA's Support to Education in Africa in the Last Two Decades: Focusing on Mathematics and Science Education." Journal of International Cooperation in Education, 19 (2): 35–53.

Mbiti, I., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." Quarterly Journal of Economics, 134 (3): 1627–1673. https://doi.org/10.1093/qje/qjz010.

Milner, H. R. IV, 2013. "Policy Reforms and De-Professionalization of Teaching." National Education Policy Center.

Muralidharan, K., and V. Sundararaman. 2011. "Teacher Performance Pay: Experimental evidence from India." Journal of political Economy, 119(1), 39-77.

Piper, B., and M. Korda. 2011. "Early Grade Reading Assessment (EGRA) Plus: Liberia: Program Evaluation Report." RTI International. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=283

Piper, B., S. S. Zuilkowski, M. Dubeck, E. Jepkemei, and S. J. King. 2018. "Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides." World Development, 106: 324–336.

Popova, A., D. K. Evans, M. E Breeding, and V. Arancibia. 2022. "Teacher Professional Development Around the World: The Gap between Evidence and Practice." World Bank Research Observer, 37 (1): 107–136.

Praetorius, A. K., E. Klieme, B. Herbert, and P. Pinger. 2018. "Generic Dimensions of Teaching Quality: The German Framework of Three Basic Dimensions." Zdm Mathematics Education 50: 407–426.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. “Teachers, Schools, and Academic Achievement.” Econometrica, 73 (2): 417–458.

Rockoff, J. E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” American Economic Review, 94 (2): 247–252.

Sabarwal, S., D. Evans, and A. Marshak. 2014. “The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone.” World Bank Policy Research Working Paper No.7021. World Bank.

Snilstveit, B., J. Stevenson, D. Phillips, M. Vojtkova, E. Gallagher, T. Schmidt, H. Jobse, M. Geelen, M. Pastorello, and J. Eyers. 2015. “Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle-Income Countries: A Systematic Review.” 3ie Systematic Review 24. International Initiative for Impact Evaluation (3ie)

Strengthening of Mathematics and Science Education (SMASE) in Africa (SMASE-Africa). 2025. “Strengthening of Mathematics and Science Education in Africa.” https://www.smase-africa.org/.

UNESCO Institute for Statistics. 2017. “More Than One-half of Children and Adolescents are not Learning Worldwide.” UIS Fact Sheet No. 46. UNESCO. https://uis.unesco.org/sites/default/files/documents/fs46-more-than-half-children-not-learning-en-2017.pdf.

UNESCO Institute for Statistics. 2025. UIS Data Browser. Retrieved from http://data.uis.unesco.org

World Bank. (2018). “World Development Report 2018: Learning to Realize Education’s Promise.” World Bank. https://doi.org/10.1596/978-1-4648-1096-1

Figure A1： Lesson Observation Checklist

		No	Criterion	1	2	3	4	n/a	Evidence
Lesson plan		1	Lesson plan has objectives of the lesson that are appropriate for student learning based on competencies in respective syllabus.						
Lesson delivery	Lesson progression	2	Teacher conducts appropriate starter activities to arouse interest of students toward the lesson.						
		3	Teacher provides students with a key question/problem/issue to think over.						
		4	Teacher provides students with activities (with T&L materials where necessary) that are appropriate and adequate to meet the lesson objectives (including improvised materials).						
		5	Teacher facilitates the class to make a summary reflecting their learning.						
	Student partici-pation	6	Students answer teacher's questions actively.						
		7	The lesson is student-centred.						
	Basic knowledge and skills of lesson delivery	8	Teacher uses chalkboard effectively.						
		9	Teacher deals with students' questions properly.						
		10	Teacher deals with students' responses properly.						
		11	Teacher gives clear instruction/necessary support to students and supervises learning activities.						
		12	Teacher uses time effectively and makes appropriate pacing of lessons for student learning.						
		13	Teacher explains concepts and ideas using appropriate language and techniques to make student easily understand the lesson.						
		14	Teacher presents subject matter contents correctly.						

Abstract (in Japanese)

要　約

教員の質は児童・生徒の学力に非常に重要な影響を与えるが、低・中所得国において政府が教員の質をどう改善できるかは明らかになっていない。この研究では、ランダム化比較試験を用いて、エチオピア政府が実施した算数・物理教員の職能開発プログラムを評価した。また、教員研修の補完的な介入として、教室での学習を定着させるためのワークブックの効果を検証した。教員研修の結果、教員の教授法や教科知識は向上したが、児童のテストスコアは短期間では向上しなかった。一方、児童が補習授業において教員の補助のもとワークブックを活用したところ、児童のテストスコアも向上した。このプログラムでは教員に金銭的なインセンティブを与えなかったが、教員はプログラムの実施後も各自の学校で努力を継続し、教授法や教科知識へのインパクトが強化された。

キーワード：　教育政策、学力、教員研修、職能開発、ランダム化比較試験