Studies on Effective Interventions to Improve Math Learning in Low- and Lower-Middle-Income Countries

Takao Maruyama

A Dissertation
Submitted to the Graduate School of Hitotsubashi University
for the Degree of Doctor of Philosophy in Economics

February 2022

Contents

List of Abbreviations

COVID-19	Coronavirus Disease 2019
DD	Difference in Differences
EFA-FTI	Education for All: Fast Track Initiative
EPT	Ecole pour Tous
ESMATE	Proyecto de Mejoramiento de Aprendizajes en Matemática en Educación Básica y Educación Media
GDP	Gross Domestic Product
IRT	Item Response Theory
JICA	Japan International Cooperation Agency
MCC	Millennium Challenge Corporation
MDGs	Millennium Development Goals
NGO	Non-Governmental Organization
PASEC	Programme d'analyse des systèmes éducatifs de la CONFEMEN
PISA	Programme for International Student Assessment
PMAQ	Paquet Minimum Axé sur la Qualité
RCT	Randomized Controlled Trial
SDGs	Sustainable Development Goals
SI	Sequential Ignorability
SMC	School Management Committee
TaRL	Teaching at the Right Level
TIMSS	Trends in International Mathematics and Science Study
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNDP	United Nations Development Programme
USAID	United States Agency for International Development

Chapter 1. Introduction: Background and Objectives of the Dissertation

1.1 A priority issue in international educational development: The "learning crisis"

In 1990, the international community declared basic education as a priority area for international development (UNESCO 1990). Basic education is a human right for all and contributes to socio-economic development (UNESCO, 1990). In the 1990s, development aid focused on poverty eradication and human development. Human development is a process of widening people's choices and levels of well-being (UNDP, 1990), and basic education builds the foundation for human development (UNESCO, 1990). Psacharopoulos (1985; 1994) demonstrated that educational attainment was associated with the earnings of individuals, and the rate of return was especially high for primary education. The studies supported the increase in development aid for basic education. Whereas developing countries realized progress toward the universalization of access to basic education, problems in access to, quality of, and equity in basic education remained (UNESCO, 2000). In this context, the Millennium Development Goals (MDGs) and Dakar Framework for Action were agreed upon, which included the goals and targets pertaining to basic education that were to be achieved by 2015. The financing mechanism, called "Education for All: Fast Track Initiative (EFA-FTI)," was also launched to support developing countries in achieving goals pertaining to basic education (Lister et al., 2010).

With the efforts of the international community, access to primary education significantly expanded in developing countries. The net enrollment rate in primary education in low-income countries improved from 60 percent in 1999 to 83 percent in 2012 (UNESCO, 2015). In 1999, access to primary education in sub-Saharan Africa lagged far behind the rest of the world, but the region significantly improved the net enrollment rate from 59 percent in 1999 to 79 percent in 2012 (UNESCO, 2015). Despite the remarkable progress in access to primary education, the quality stagnated in developing countries. Pritchett and Beatty (2015) presented the low levels of learning in reading and mathematics of primary-school-age children in India, Kenya, Pakistan, Tanzania, and

Uganda, and revealed a profound mismatch between the curricula and actual learning levels. Altinok et al. (2014) linked the student test scores of 103 countries in different international or regional assessments in basic education. The study revealed that average learning outcomes in developing countries lagged far behind those in developed countries, and a large percentage of students in developing countries did not reach the minimum threshold level of learning that they should have mastered in primary education (Altinok et al., 2014). UNESCO (2014) called the situation wherein a large number of children spent years in school without gaining foundational skills, the "learning crisis."

Considering the alarming situation of the low quality of education in developing countries, the international community set the target of ensuring all girls and boys completing free, equitable, and quality primary and secondary education in the Sustainable Development Goal 4 (SDG 4) (United Nations General Assembly, 2015). The key indicator for the target is the proportion of children in primary and lower secondary levels of education who achieve minimum proficiency levels in reading and mathematics (UNESCO, 2016).[1] UNESCO (2017) estimated that over 600 million children of primary and lower secondary school age were not reaching the minimum proficiency levels in reading and mathematics in low- and lower-middle-income countries. The situation was especially acute in sub-Saharan Africa. Over 80 percent of children of primary and lower secondary school age were not mastering the basics of reading and mathematics (UNESCO, 2017). The situation in other regions was also alarming. In Latin America and the Caribbean region, approximately half of the children were not reaching the minimum proficiency level in mathematics (UNESCO, 2017). In the region, access to primary education was almost universal in the early 2000s, but the quality of primary education

[1] The working definition of the minimum proficiency levels in reading and mathematics refers to the assessment frameworks of the international or regional assessment in basic education, including the Trends in International Mathematics and Science Study (TIMSS) and *Le Programme d'analyse des systèmes éducatifs de la CONFEMEN* (PASEC) (UNESCO, 2021). For example, PASEC, the regional standard assessment for French-speaking African countries, evaluates student learning in reading and mathematics in primary education by five levels: from "below level one" to "level four." In the assessment framework, the minimum proficiency level in mathematics in primary education is level two, at which level students of the last grade of primary must have mastered the basics of arithmetic, measurement, and geometry (PASEC, 2014; 2019). In reading, those who are able to combine two pieces of explicit information from a document or carry out simple inference from a text are assessed to be above the minimum proficiency level (PASEC, 2014; 2019).

stagnated over the years (Angrist, 2021). The recent long-term closure of schools due to the COVID-19 pandemic has further worsened the global learning crisis (UNESCO, 2021). Addressing the learning crisis in basic education is a priority issue in international educational development.

Whereas global disparity in learning outcomes is an issue in international educational development, domestic disparity by economic status exists in developing countries. Akmal and Pritchett (2021) constructed a household wealth index from information on household asset in India, Kenya, Pakistan, Tanzania, and Uganda and presented the percentage of children who mastered literacy, and numeracy by the index. The bottom 40 percent of children in the wealth index were significantly less likely to master foundational skills than the top 20 percent. The gap in learning outcomes by household economic status will further generate economic disparity in the next generation.

1.2 The importance of improvement in math learning for human development and economic growth

In mathematics in primary education, children learn numbers and four basic operations, which are essential skills for daily life. Children also study quantity and measurement, and geometry, which are closely related to daily contexts as well. The knowledge and skills that children acquire in mathematics in primary education are foundational. Empirical studies suggest that basic cognitive skills help people build technical or professional skills later in life. Hanushek et al. (2015) analyzed microdata of developed countries, and identified the positive impact of foundational skills, literacy, and numeracy on individuals' earnings. One standard deviation increase in numeracy skills was associated with the increase in hourly wages for prime age workers by 18 percentage point on average (Hanushek et al., 2015). Moreover, Foster and Rosenzweig (1996) and Rosenzweig (2010) argued that farmers with basic cognitive skills were likely to adopt new technologies.

By learning mathematics in basic education, children develop logical reasoning and problem-solving skills. In a knowledge-based society, it becomes important to develop skills of applying knowledge to solve problems in various contexts. Thus, OECD (2003) developed a concept of mathematical literacy for international standard assessment called the Programme for International Student Assessment (PISA). Mathematical literacy is the capacity to formulate problems in various contexts using knowledge of mathematics, solve the math problems, and interpret the results (OECD, 2003). Children use mathematics to describe, explain and predict phenomena. Mathematics is also a foundation for children to study other subjects, such as science and information technology (OECD, 2021).

Improving math learning contributes to a country's economic growth. Studies in the economics of education demonstrated that the improvement in learning outcomes rather than the number of years of schooling was a source of sustained economic growth (Hanushek and Woessman, 2012; 2015). Through a country-level panel data analysis, Hanushek and Woessman (2015) showed that one standard deviation increase in the international or regional assessment in mathematics was associated with around two percentage point increase in annual growth rate of the GDP per capita, even after controlling for the number of schooling years. In the context of the severe learning crisis in low- and lower-middle-income countries, the improvement of math learning in basic education is a important development agenda for human development and economic growth.

1.3 Objectives and methods

To address the learning crisis in low- and lower-middle-income countries, development agencies conducted various interventions using different approaches (Snilstveit et al., 2016). The interventions were provided for different targets, including children, households, systems, schools, and teachers (Snilstveit et al., 2016). Based on several systematic reviews, Evans and Popova (2016) argued that whereas pedagogy tailoring methods and contents of teaching to student learning levels was effective in improving learning, the question remained as how it was possible to realize such teaching practices in the context of developing countries. The capacities of the local educational administration and teachers are often limited in developing countries (Bashir et al., 2018; World Bank, 2018).

World Bank (2018) raised four immediate determinants of the learning crisis: (i) unmotivated and unskilled teachers, (ii) lack of readiness of learners in early grades, (iii) lack of effective school management, and (iv) insufficient or untimely provision of school inputs such as teaching and learning material. Any of the four determinants can become a bottleneck for interventions to improvement learning. For example, the magnitude of the impact of the provision of learning material in mathematics on student learning could become smaller by one of the first three factors mentioned above (World Bank, 2018). To address these multiple factors, it is necessary to develop a package of interventions to improve math learning. The package should be scalable in the context of developing countries that have local educational administrations and human resources with limited capacity.

While numeracy is an essential foundational skill that children learn in mathematics, in basic education, they continue their math learning in various other domains such as quantity and measurement, and develop applying and reasoning skills. Development aid in basic education should cover various domains in math curricula, and help students develop applying and reasoning skills. In math teaching and learning, textbooks are important intermediates that link curricula, teachers, and students. Since textbooks cover curricula and define teaching and learning contents, the development of quality textbooks could bring about better math learning in developing countries. While the quality of daily math lessons could be improved by ensuring the quality of textbooks, the learning levels of children in the same grades are often heterogeneous and do not reach the level expected in the curricula (Pritchett and Beatty, 2015). Parents should support their children in keeping pace with the school curricula; however, parents face challenges for various reasons, such as lack of information on student's academic achievements or low educational attainment (Dizon-Ross, 2019). Thus, it is necessary to complement the intervention strategy that addresses the quality of math textbook by using other strategies that support weaker children in catching up with math learning.

This dissertation presents evidence pertaining to the two different packages of interventions that aim to improve math learning in basic education. Both packages were developed in technical cooperation by the Japan International Cooperation Agency (JICA), a bilateral aid agency of the government of Japan. The agency conducted a series of projects on technical cooperation in basic education in Central America and West Africa since the 2000s. One of the packages was developed in El Salvador, which was composed of distribution of learner-friendly math textbooks and other related components, such as introductory teacher training (JICA, 2019). The math textbooks covered from the first grade in primary to the last grade in secondary education. Subject contents of mathematics were carefully subdivided considering the student assessment, and the contents were sequenced in the textbook to assure small-step learning by students. The textbooks were also designed to facilitate lessons in which students engaged in learning mainly through problem solving.

The other package of interventions was developed in Niger, which was composed of a series of training for the school management committee (SMC), and a provision of math workbooks accompanied with a training for teachers and community volunteers (Hara et al., 2020). SMC was composed of the school principal and representatives from teachers, parents, and community members. After the training for SMC executive members, the SMC organized a community general assembly where it shared a summary of math assessment results with parents, teachers, and community members. The committee facilitated discussions on the development of joint action plans integrating extra-curricular remedial activities using math workbooks. Math workbooks were self-learning materials for students, which were distributed for schools by the Ministry of Education. Students received workbooks that matched their learning levels and studied math at their own pace. Teachers and community volunteers took on the role of activity facilitators of checking students' answers and giving them instructions. The package was scaled up by the Ministry of Education in Niger to cover around 3,500 schools in the Tillaberi region in 2018.

This dissertation utilizes data collected by the JICA supported projects in El Salvador and Niger, and rigorously estimates the impact of the package of interventions on student math learning. In El Salvador, the Ministry of Education and JICA jointly conducted a randomized controlled trial

(RCT) to measure the impact of the package of interventions on student math learning. The experiment targeted 2nd grade in primary education, and 7th grade in lower secondary education. In a series of survey, same students were tracked for two years, 2018 and 2019.

In Niger, JICA supported project conducted a series of survey for sampled schools. Originally, the package of interventions targeted all the six grades in primary education. But due to budget constraints, the distribution of math workbooks was limited to students from the 1st to 4th grades only. Focusing on the difference in the intervention, this dissertation estimates the causal impact of the package of interventions on 4th grade student math learning in comparison with 5th grade students using the difference in differences (DD) strategy.

This dissertation contributes to the literature on international educational development by demonstrating evidence pertaining to the two complementary packages of interventions to improve math learning, and exploring the external validities for other contexts. The package of interventions developed in El Salvador mainly targeted the improvement of math learning in regular classes. On the other hand, the package of interventions developed in Niger focused on extra-curricular remedial activities in math. Whereas the former package of interventions covered math curricula, the latter focused on basics in mathematics, numbers and four basic operations.

A popular approach to exploring the external validity of evidence is the systematic review that synthesizes evidence from different countries; however, the policy implications from the reviews are often too broad to apply to specific contexts (Williams, 2020). Therefore, instead of attempting a systematic review, this dissertation uses a method called "mechanism mapping" proposed by Williams (2020) so that the external validities of the evidence from El Salvador and Niger can be explored. The method begins by laying out a theory of change of the intervention. Subsequently, important contextual assumptions related to each step in the causal chain are raised. Finally, actual contextual situations are laid out corresponding to the contextual assumptions. The differences between the contextual assumptions and actual situations indicate the parts in a theory of change, which should be examined and adapted to the actual context. The evidence pertaining to the two packages and examination of the external validities jointly suggests hints to other developing countries to improve student math learning in basic education.

This dissertation also contributes to the literature on evidence use in development aid for education through a case study of an organization that actively utilized evidence to improve educational aid. While it depends upon development agencies to utilize evidence to improve strategy and practice, the literature mainly supposes one-way communication from researchers to development agencies, rather than an interactive process between the two entities (Langer et al., 2015). It does not focus on organizational aspects such as the initiatives and management of development agencies. A prominent example of an organization that utilized evidence to improve operations in educational aid is the Indian NGO "Pratham" (Banerjee et al., 2017; Duflo, 2020). Since the early 2000s, Pratham has conducted a series of rigorous evaluations in collaboration with researchers. As the case of evidence use by Pratham is longitudinal and includes several examples of impact evaluation, it describes how the organization actively utilized evidence to improve operations.

1.4 Outline of this dissertation

This dissertation is composed of five chapters. The second and third chapters present the results of a RCT in El Salvador. The second chapter focuses on the impact of the package of interventions on student math learning in primary education, and the third chapter demonstrates the results of the analysis at the lower secondary level of education. The fourth chapter estimates the impact of the package of interventions on student math learning in Niger. The fifth chapter explores the external validities of evidence from El Salvador and Niger, using the method of "mechanism mapping." The sixth chapter is on the case study of evidence use by the Indian NGO "Pratham," and draws lessons for other development agencies on evidence use.

The following are the summaries of Chapter two through six. Chapter 2 evaluates the effectiveness of a package of interventions, including the distribution of textbooks that are carefully designed to improve math learning, using data collected in the RCT in El Salvador. To capture the progress in learning over two years, the chapter linked math test scores at the end of the primary 2nd grade and the 3rd grade, using the item response theory (IRT). The average one-year impact of the package on math learning is estimated 0.49 standard deviations of the IRT scores. The average accumulated impact of the first-year interventions in the following year is 0.13 standard deviations of the IRT scores. The package of interventions improved math learning, and the impact persisted even after schools in the control group also received the package of interventions in the following year.

The Ministry of Education of El Salvador scaled up the package of interventions nationwide for grades 7 through 9 (the lower secondary education level), but the ministry could not cover several parts of the program, including the distribution of student math workbook and diagnostic math test, because of budget constraints. Chapter 3 investigates the impact of the components, which were not covered by the Ministry of Education, in a package of interventions on math learning at the lower secondary level in El Salvador. The study also linked test scores at different rounds of the survey by the item response theory (IRT). In year 1 of the research, the additional components had a positive impact for students with lower household economic status, and the average impact is estimated at 0.17 standard deviations. However, the impact did not persist in the following year. Causal mediation analysis is employed to unpack the causal path on the first-year impact.

Chapter 4 investigates the impact of a package of interventions that included training for SMCs and the distribution of math workbooks on math learning in Niger. The package was scaled up by the ministry of education for approximately 310,000 students in the 1st to 4th grades in approximately 3,500 public primary schools. The scaled-up interventions aimed to help the students improve their math learning through extra-curricular remedial activities that they participated in over a three-month period. Due to budget constraints, the distribution of math workbooks was limited to students from the 1st to 4th grades among the six grades in primary education. Focusing on the difference in the intervention between the 4th and 5th grade students, the study investigates the impact of the package of interventions on student math learning using three-round survey data. The study finds that the average impact of the interventions on math learning outcomes is estimated to be 0.36 to

0.38 standard deviations. The impact is larger for students with lower baseline scores.

Chapter 5 explores the applicability of the packages of interventions in El Salvador and Niger to other low- and lower-middle-income countries, using the analysis framework called "mechanism mapping" proposed by Williams (2020). The mapping exercise clarifies the important contextual assumptions to be reviewed when applied for other contexts. For example, the development of math textbooks depends on the contextual assumptions from the institutional aspect such as math curricula, cultural aspect such as mother tongue and instructional language, and organizational aspect such as the capacity for printing/procuring and delivering textbooks for schools. The fifth chapter discusses the applicability of the packages of interventions developed in El Salvador and Niger to other countries, and possible measures to enhance external validities.

Chapter 5 also presents a case study of evidence use by the Indian NGO "Pratham" in the appendix. In a series of impact evaluations, Pratham identified the research agenda that addressed the gap between the mission and actual level of performance of the organization. Since the research agenda originated from the strategic concerns of Pratham, the results of the experiments were highly relevant to the strategy and operations of the organization. The leading management in Pratham played a central role in promoting and coordinating the experiments. With the active involvement of the leading management, impact evaluations were embedded within the operational and decision-making structure of development agencies, by which evidence was fed back well into the strategy and operations of the organization. Based on an equal partnership between Pratham and researchers, knowledge and expertise were exchanged in the process of experiments effectively, which made the evaluations productive. The lessons from the case study of Pratham provide insights for bilateral or multilateral development agencies to utilize evidence to improve and expand their operations in education aid.

1.5 Issues for future research

1.5.1 Study of the intervention strategy for improving math learning

This dissertation studies the package of interventions for improving student math learning in El Salvador and Niger. The package of interventions in El Salvador, including the distribution of textbooks covered math curricula, and the math workbooks in Niger focused on numbers and four basic operations. Whereas the package of interventions in El Salvador improved math learning in primary 2nd grade as reported in Chapter 2, the correct response rate to the items posed in text in the treatment group remained low at the end of the school year. For example, approximately half of the students in primary 2nd grade in the treatment group correctly answered a simple division problem; however, the correct response rate to the item posed in text using simple division was approximately 15 percent. Applying and reasoning skills in mathematics may be enhanced by improvements in basic reading and reading comprehension. In future research on the interventions in math learning, it would be interesting to add a component to improve basic reading and reading comprehension.

As presented in Chapter 3, most of the 1st-grade students in lower secondary education in El Salvador did not master the basics of mathematics in primary school, including four basic operations. In years 1 and 2 in the research, the students struggled with content in mathematics that became theoretical and abstract in lower secondary education. In the math tests conducted in the research, the correct response rates to the math items of the simple equation were very low. The results indicate that it is essential to provide support for students to catch up on the basics of math that should have been mastered at primary school. However, it is also necessary to examine whether the students in El Salvador could make progress in their math learning at the lower secondary level with the textbooks developed by the JICA-supported project after they catch up with the basics in primary education. In the research, a series of survey was administered that consisted of a written math test that measured learning outcomes. However, to obtain insights for the revision of the textbooks, it would be helpful to conduct a survey on students' math learning process, for example, by observing how they solve math problems in math lessons, conducting student error analysis on math items, or organizing student interviews.

1.5.2 Study of inequality in math learning outcomes

In the context of the learning crisis, through it is important to improve math learning outcomes on average in low- and lower-middle-income countries, it is necessary to pay attention to the disparity in math learning outcomes. Chapters 2 and 3 of this dissertation investigate the heterogeneous impact on math learning outcomes by household economic status. Chapter 4 also investigates heterogeneous impact on math learning outcomes by gender. Several inequality measures have been proposed in recent studies. Ferreira and Gignoux (2014) measured the inequality of educational opportunity, using the PISA dataset. The inequality of educational opportunity was measured by the share of the variance in test scores, which was explained by the predetermined characteristics of students. Rodriguez-Segura et al. (2021) presented several inequality measures in learning outcomes, such as the Gini coefficient. Such measures can be employed to deepen the analysis of impact of interventions in future research

1.5.3 Study of social learning among teachers

In El Salvador, the experiment targeted 2nd grade of the primary school, and the survey tracked them for two years. The survey also tracked the same teachers in end-line (year 1 of the research) and follow-up surveys (year 2 of the research), and observed math lessons. Chapter 2 of this dissertation investigates the accumulated impact of the first-year intervention in year 2 of the research. Whereas the package of interventions had an accumulated impact on math learning for the primary students in year 2, the regression results showed that it did not have an accumulated impact on teaching practices. The textbooks developed in the JICA-supported project were distributed to primary schools nationwide in year 2 of this research. Since teachers use the same textbook, they can share their findings and experiences with their colleagues, and improve their practices.

Jackson and Bruegmann (2009) investigated the human capital spillover for teachers, using a longitudinal data set of student test scores linked to teacher characteristics in the United States. They showed that students obtained larger scores gains when their teachers experienced improvements in the observable characteristics of their colleagues (Jackson and Bruegmann, 2009). Social learning has been studied in other fields, such as the agriculture in developing countries. The diffusion of technology through social learning among farmers has been investigated (Conley et al., 2010; Takahashi et al., 2019). In educational development, it would be worth conducting research on social learning among teachers, referring to the methods developed in agricultural studies.

1.5.4 Study of the external validity of evidence, and evidence use by Pratham

A popular method for exploring the external validity of evidence is systematic review, which collects evidence on a specific domain, examines its quality, and synthesizes it (White and Waddington, 2012). Chapter 5 of this dissertation explores the external validities of evidence from El Salvador and from Niger, using the method of "mechanism mapping" proposed by Williams (2020). Mechanism mapping explicitly lays out a theory of interventions and the contextual assumptions. Whereas the method of mechanism mapping was developed to explore the external validity of a specific context, the results of mechanism mapping can be used to enhance the quality of systematic reviews in educational development. Systematic reviews can more deeply examine the mechanism of the intervention, using the results of the mapping exercise.

In the practice of aid, development agencies have been applying an approach developed in a country for other contexts. For example, JICA developed a package of interventions in Niger and applied the package to other countries such as Madagascar (Maruyama et al., 2021). In the process, the project team in Madagascar has adapted the package developed in Niger into the local context of Madagascar. It is worth documenting the experience to deepen the understanding of the external validity of evidence and obtain insights on the practice in development aid.

Chapter 6 focuses on a series of experiments conducted by Pratham in collaboration with researchers. Rigorous evaluations clearly revealed the impacts of interventions, which helped Pratham develop and establish the scaling-up strategies of their pedagogy for the leading program “Read India.” The program was estimated to reach around 16.6 million in 21 states in 2018 (Pratham Education Foundation, 2019). When Pratham mainstreamed the strategies and methods developed and established through experiments, it shared them widely in the organization. Utilizing evidence in educational aid is not limited to the stages of research but includes human resource development and knowledge sharing. It would be worth continuing research on how Pratham widely disseminated strategies and methods established through research among its staff.

Reference

Akmal, Maryam, Lant Pritchett. 2020. “Learning equity requires more than equality: Learning goals and achievement gaps between the rich and the poor in five developing countries.” *International Journal of Educational Development*, 82 (2021) 102350.

Altinok, Nadir, Claude Diebolt and Jean-Luc Demeulemeester. 2014. “A new international database on education quality: 1965–2010.” *Applied Economics*, 46:11: 1212-1247.

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

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobini Mukerji, Marc Shotland, and Michael Walton. 2017. “From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application.” *Journal of Economic Perspectives*, Volume 31, No. 4, Fall 2017: 73-102.

Bashir, Sajitha, Marlaine Lockheed, Elizabeth Ninan, and Jee-Peng Tan. 2018. *Facing Forward: Schooling for Learning in Africa. Africa Development Forum Series*. Washington DC: World Bank.

Conley, Timothy G., and Christopher R. Udry. 2010. “Learning about a New Technology: Pineapple in Ghana.” *American Economic Review*, 2010, 100:1: 35-69.

Dizon-Ross, Rebecca. 2019. “Parents’ Beliefs about Their Children’s Academic Ability: Implications for Educational Investments.” *American Economic Review*, 109(8): 2728-65.

Duflo, Esther. 2020. “Field Experiments and the Practice of Policy.” *American Economic Review*, 110 (7): 1952-1973.

Evans, David K., and Anna 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.

Ferreira, Francisco H. G., and Jérémie Gignoux. 2014. “The Measurement of Educational Inequality: Achievement and Opportunity.” *The World Bank Economic Review*, 2014, Vol. 28, No. 2: 210-246

Foster, Andrew D and Rosenzweig, Mark R, 1996. “Technical Change and Human-Capital Returns and Investments: Evidence from the Green Revolution.” *American Economic Review*, vol. 86(4): 931-953.

Hanushek, Eric A., and Ludger Woessmann. 2012. “Schooling, educational achievement, and the Latin American growth puzzle.” *Journal of Development Economics*, 99 (2): 497-512.

Hanushek, Eric A., and Ludger Woessmann. 2015. *The Knowledge Capital of Nations: Education and the Economics of Growth*. Cambridge: The MIT Press.

Hanushek, Eric A., and Ludger Woessmann. 2015. “Returns to skills around the world: Evidence from PIAAC.” European Economic Review, vol. 73, issue C: 103-130.

Hara, Masahiro, Takao Maruyama, Akiko Kageyama, and Nobuhiro Kunieda. 2020. “Quality Learning through Community-wide Collaboration: A Methodology to Overcome the “Learning Crisis” in Niger.” In M. Nishimura (Ed.), Community participation with schools in developing countries: Towards equitable and inclusive basic education for all. Routledge Research in Educational Equality and Diversity Series. New York & Oxon: Routledge.

Jackson, C. Kirabo, and Elias Bruegmann. 2009. “Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers.” *American Economic Journal: Applied Economics*, 1:4, 85–108.

JICA 2019. *Project Completion Report of Project for the Improvement of Mathematics Teaching in Primary and Secondary Education in El Salvador. (in Japanese)*.

Langer, Laurenz, Ruth Stewart, Yvonne Erasmus, and Thea de Wet. 2015. "Walking the last mile on the long road to evidence-informed development: building capacity to use research evidence." *Journal of Development Effectiveness*, 1095783.

Lister, Stephen. Manos Antoninis, Felicity Binns, Roy Carr-Hill, Mailan Chiche, Catherine Dom, Keith Hinchliffe, Fred Merttens, Georgina Rawle, Abby Riddell, Anthea Sims Williams, Muriel Visser-Valfrey, and Eric Woods. 2010. *Mid-Term Evaluation of the EFA Fast Track Initiative Final Synthesis Report: Volume 1 – Main Report*. EFA-FTI.

Maruyama, Takao, Kengo Igei and Seiichi Kurokawa. 2021. "Community-wide support for primary students to improve learning: Empirical evidence from Madagascar." RISE Annual Conference 2021.

OECD 2003. *The PISA 2003 Assessment Framework: Mathematics, Reading, Science, and Problem-solving Knowledge and Skills*. OECD.

OECD 2021. PISA 2021 Mathematics Framework. Retrieved from https://pisa2021-maths.oecd.org/

Pratham Education Foundation 2019. *Annual reports 2018-19*. Pratham Education Foundation.

Pritchett, Lant, and Amanda Beatty. 2015. "Slow Down, You Are Going Too Fast: Matching Curricula to Student Skill level." *International Journal of Educational Development*, 40: 276-288.

Psacharopoulos, George. 1985. "Returns to Education: A Further International Update and Implications." *Journal of Human Resources*, 20(4): 583-604.

Psacharopoulos, George. 1994. "Returns to Investment in Education: A Global Update." *World Development*, 22(9): 1325-43.

Rodriguez-Segura, Daniel, Cole Campton, Luis Crouch, and Timothy S. Slade. 2021. "Looking beyond changes in averages in evaluating foundational learning: Some inequality measures." *International Journal of Educational Development*, 84 (2021) 102411.

Rosenzweig, Mark R. 2010. “Microeconomic Approaches to Development: Schooling, Learning, and Growth.” *Journal of Economic Perspectives*, 24 (3): 81-96.

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt and Emmanuel Jimenez. 2016. *The Impact of education programmes on learning and school participation in low- and middle- income countries*. London: International Initiative for Impact Evaluation.

Takahashi, Kazushi, Yukichi Mano, and Keijiro Otsuka. 2019. “Learning from experts and peer farmers about rice production: Experimental evidence from Cote d'Ivoire.” *World Development*, Vol. 122, Oct. 2019: 157-169.

UNESCO 1990. *World Declaration on Education For All Meeting Basic Learning Needs*. World Conference on Education for All. Jomtien, Thailand, 5-9 March 1990, UNESCO, New York.

UNESCO 2000. *Dakar Framework for Action: Education for All. Meeting Our Collective Commitments*. World Forum on Education, Dakar, Senegal, 26-28 April 2000, UNESCO, Paris.

UNESCO 2014. *EFA Global Monitoring Report 2013/4: Teaching and Learning: Achieving Quality of All*. Paris: UNESCO Publishing.

UNESCO 2015. *EFA Global Monitoring Report 2015: Education for All 2000-2015: Achievement and Challenges*. Paris: UNESCO Publishing.

UNESCO 2017. “More than One-half of Children and Adolescents Are Not Learning Worldwide.” *Fact Sheet No. 46*. (UIS/FS/2017/ED/46).

UNESCO 2021. *Pandemic-related disruptions to schooling and impacts on learning proficiency indicators: A focus on the early grades*. Montreal: UNESCO Institute for Statistics.

UNDP 1990. Human Development Report 1990. New York: Oxford Press.

United Nations General Assembly. 2015. “Transforming our world: the 2030 Agenda for Sustainable Development.” October 21, 2015, A/RES/70/1.

White, Howard and Hugh Waddington. 2012. “Why do we care about evidence synthesis? An introduction to the special issue on systematic reviews." *Journal of Development Effectiveness*, 4 (3): 351-358

Williams, Martin, J. 2020. "External Validity and Policy Adaptation: From Impact Evaluation to Policy Design." *The World Bank Research Observer*, Vol. 35 (2): 158–191.

World Bank 2018. *World Development Report 2018: Learning to Realize Education's Promise*. Washington DC: World Bank.

Chapter 2. Developing Textbooks to Improve Student Math Learning: Empirical Evidence from El Salvador[2]

2.1 Introduction

In teaching and learning practices, textbooks are an important intermediate that links curriculum, teachers, and students since they describe the content and methodology of teaching and learning. While textbooks were regarded as one of the most important inputs to significantly impact student learning up until 2000 (World Bank 2001), recent evidence shows that the magnitude of the impact of the distribution of textbooks on student learning is small (Glewwe 2013). Distribution of textbook only improved learning among students with high baseline scores (Glewwe et al. 2009), and those with high socioeconomic status (Kuecken and Valfort 2013). The results indicate that the expansion of school enrollment in the 2000s brought a mismatch between existing textbooks and student learning levels.

Snilstveit et al. (2016) argue that it is necessary to address several challenges such as inadequately trained teachers and a lack of appropriate materials, curricula, and instructional approaches for improving learning. If a textbook is carefully designed based upon student learning levels, it can address those challenges in conjunction with other types of interventions. For example, since textbook structures daily lessons, it can change teaching practices of teachers to better support student learning. Lesson observations and feedback for teachers can help teachers enhance their teaching practices (Bruns et al. 2018). A package of interventions including textbook development can cover content in math curricula in primary education.

To fill in this lack of understanding regarding effective teaching of math through textbook development, a novel package of interventions has been developed in El Salvador, which this chapter analyzes. In 2015, the Ministry of Education of the country launched the "Project for the Improvement of Mathematics Teaching in Primary and Secondary Education" ("Proyecto de Mejoramiento de

[2] This chapter is developed from the following paper and the revised version,
Maruyama, Takao, and Takashi Kurosaki. "Developing textbooks to improve student math learning: Empirical evidence from El Salvador." *JICA Ogata RI Working Paper*, No. 217.

Aprendizajes en Matemática en Educación Básica y Educación Media" in Spanish, hereinafter "ESMATE"), with technical cooperation from Japan International Cooperation Agency (JICA). In the traditional approach to teaching mathematics in El Salvador, teachers explained how to solve math problems and posed example problems. Students listened to the explanations and took notes, but most of them merely copied the answer without understanding how it was derived. Teachers did not pay attention to how much students learned (JICA 2019). This situation is common in Latin American countries. Through lesson observations in several countries in the region, Bruns and Luque (2015) found that teachers failed to keep students engaged in learning, which resulted in poor student learning results.

The textbooks developed by the ESMATE project, hereinafter referred to as "ESMATE textbooks," intended to change such teaching and learning practices. Subject contents were carefully subdivided considering the student assessment conducted by the project, and the contents were sequenced in the textbook to assure small-step learning by students. The textbooks were designed to facilitate lessons in which students engage in learning mainly through problem-solving work.[3] Then, the ESMATE project combined the distribution of textbooks with the other types of interventions including introductory teacher training and lesson observations by the school principal. This study evaluates the impact of the package of interventions on primary 2nd grade students' math learning outcomes using a randomized controlled trial.[4][5] This experiment continued for two years and tracked the same students through three rounds of surveys in 2018 (baseline, end-line) and 2019 (follow-up). In the second year, both treatment and control groups received interventions that included grade 3 textbooks.

[3] Textbooks, workbooks, and teachers' guidebooks developed by the project are posted on the following website by the Ministry of Education, Science and Technology in El Salvador: https://www.mined.gob.sv/materiales-educativos/item/1014902-esmate.html Sample pages of the textbook and teacher's guidebook are shown in Appendix 1 of this chapter.

[4] The randomized controlled trial was conducted under an agreement between the Ministry of Education in El Salvador and JICA in June and October 2018. The survey and database construction were done by Koei Research & Consulting Inc., under a contract with JICA, in collaboration with the Ministry of Education in El Salvador. Hitotsubashi University Research Ethics Examination Committee reviewed the research plan. All the data used for this chapter was provided by JICA.

[5] The education system in El Salvador is composed of pre-school (age 4 to 6), primary (grade 1 to 6), lower secondary (grade 7 to 9), upper secondary (general: grade 10 to 11, and vocational: grade 10 to 12), and tertiary. Primary and lower secondary education are compulsory.

This chapter tracks the progress of math learning for two years, linking student test scores at the end-line survey with the scores at the follow-up survey using the item response theory. Since both groups received interventions including grade 3 textbooks in year 2, the accumulated impact of the first-year intervention on student math learning can be evaluated in the second year, relative to the impact of receiving the treatment only in the second year.[6] Most impact evaluations measure the effectiveness on student learning just after the intervention, but since students continue learning for years, it is important to track the progress and evaluate the accumulated impact of the first-year intervention on learning in subsequent years. The average one-year impact on primary 2nd grade student math learning is estimated at 0.49 standard deviations of the IRT scores. Then, the average accumulated impact of the first-year intervention in the following year is estimated to be 0.13 standard deviations of the IRT scores. Based on the first-year gains, students in the treatment group advanced their math learning in the following year. The result of this study underlines the importance of textbook development to improve student math learning.

This chapter also assesses the heterogeneity of impacts with respect to two aspects that are important in El Salvador. One of them is household economic status. In El Salvador, economic inequality is historically high. The Gini index in income was 54.5 in 1998 and gradually decreased as a result of modest economic growth within the country, reaching 43.4 in 2013 (World Bank 2020). Information on different types of household assets was collected through a student interview as part of the baseline survey, then constructed a household economic status index by the principal component analysis. The estimated value of the heterogeneous impact by the household economic status is negative but not statistically significant. As reported in the following section, the academic achievement of students is moderately correlated with their household economic status in the country. This chapter thus evaluates the heterogeneous impact with respect to baseline scores as well. The estimated value of the heterogeneous impact by the student baseline scores is positive but not statistically significant, either. While the ESMATE program improved student math learning on average, it requires complementary interventions like remedial activities for weak students to catch up

[6] This chapter does not refer to the impact as a “two-year impact” since that would refer to the impact on a treatment group that received two years of intervention compared to a control group that received no interventions over two years. In this study, the control group received an intervention in the second year.

with daily classes.

The remainder of this chapter is organized into the following four sections. Section 2.2 describes the experimentation design of this research and the package of interventions. Section 2.3 reports the impact on student math learning and the impacts on teaching practices in the section. Section 2.4 discusses the estimated impacts, and section 2.5 concludes.

2.2. Experimental Design

2.2.1 Context and the contents of intervention

While primary (grade 1 to 6) and lower secondary (grade 7 to 9) education are compulsory in El Salvador, the net enrolment rate of primary education was around 85 percent in 2017. Around seven to eight students out of ten reach the last grade of primary education (UIS 2017). In the latest international standard assessment that the country participated in, TIMSS 2007, around 80 percent of primary 4th grade students did not reach the lowest benchmark in math (Mullis et al. 2007). In this research, a baseline survey was conducted for primary 2nd grade students at the beginning of the school year in 2018. The baseline test results indicate low levels of learning. While around 80 percent of students correctly responded to a single-digit addition problem “1+3”, the percentage of students who could correctly answered a single-digit subtraction problem “5-3” was around 25 percent. For the item that asked students to count the number of 17 circles written in the test sheet, the correct response rate was around 50 percent.

The ESMATE project developed a set of mathematics textbooks, teachers’ guidebooks, and student workbooks. Subject contents are carefully subdivided considering the student assessment conducted by the project, and the contents are sequenced in the textbook to assure small-step learning by students. The ESMATE textbooks were designed to facilitate lessons in which students engage in learning mainly through solving math problems. To change the practices of teachers and enhance student math study at home with the teaching and learning materials, the ESMATE project designed a package of interventions for schools, hereinafter referred to as “ESMATE program.” The program was composed of several components: (a) distribution of ESMATE textbooks, student workbooks, and teachers’ guidebooks; (b) introductory teacher training on the use of the textbook; (c) regular lesson observations by school principals; (d) review meetings among teachers based on the result of tests; and (e) introductory training of representatives of the parent association.

Introductory training for teachers took place over two days, and included an explanation of the ESMATE textbook and the pattern of a lesson within the textbook. In the training, teachers also developed an annual math teaching plan. The plan is a simple one-page year-long calendar that defines which page of the textbook will be taught on which day. The school principals were advised to observe math lessons regularly, in total, four to five times a year, to give feedback to teachers. Representatives from parent associations participated in a one-day training that focused on the importance of study at home to improve math learning. Mutual review meetings were held in between semesters three times a year, at which teachers brought their students' math test results to review with other colleagues.[7] In El Salvador, lesson observations by school principals and mutual review meetings had been regularly conducted, so the ESMATE project tried to align the content of this existing work to improve math learning.

This research selected 2nd grade from primary education level and 7th grade from secondary education level, and evaluated the impact of the ESMATE program on math learning outcomes using a randomized controlled trial. Considering the difference in the content of interventions and educational levels of those two grades, this chapter focuses on primary education and discusses the impact on 2nd grade students.[8] The survey of this research continued for two years to investigate the accumulated impact of the first-year intervention in the following year. In 2018 (year 1), while the schools in the treatment group received the ESMATE program, there was no intervention by the ESMATE project on the control group. Then, in 2019 (year 2), the Ministry scaled up the package of interventions for all the grades of primary education in public schools across the country including the schools in the control group.[9] The schools in the treatment group also received ESMATE textbooks, teachers'

[7] Mutual review meetings were organized at the local level ("departamento") in this experiment. The meetings were facilitated by the local office of the Ministry of Education with technical support through the project.
[8] In 2018 (first year of evaluation), the Ministry distributed ESMATE textbook and teachers' guidebooks for grades 7 through 9 nationwide. The evaluation for 7th grade students mainly focused on the impact of the distribution of student workbooks. The result of the experimentation for 7th grade will be discussed in Chapter 3.
[9] The ESMATE project was finished at the end of June 2019. The Ministry integrated the activities in the policy

guidebooks, and 3rd grade student workbooks from the Ministry.

2.2.2 Sampling

Primary education in El Salvador is divided into two cycles: cycle 1 (grades 1 to 3) and cycle 2 (grades 4 to 6). Public schools in the country are called "basic education public schools" which can have preschool, primary, and lower and upper secondary levels according to the local educational needs. The lower secondary level corresponding to grades 7 to 9 comprises cycle 3. Since this research project targets 2nd grade for primary education and 7th grade for lower secondary education, the sampling frame was composed of basic education public schools offering cycles 1 and 3.[10] Geographically, there are 14 departments ("departamento" in Spanish) in the country including the capital city, San Salvador. The departments of Cabañas, La Union, San Miguel, and San Vicente in the central and eastern parts of the country were selected based on their educational statistics such as enrollment and drop-out rates. The educational situation in those four departments is close to or below national averages (Table 1).

Table 1. Educational Statistics in the four departments

	National	Cabañas	La Union	San Miguel	San Vicente
Primary net enrollment rate (2015)	86.2%	89.0%	81.2%	85.7%	85.7%
Primary repetition rate (2014)	5.8%	6.7%	5.5%	5.4%	7.7%
Primary drop-out rate (2014)	6.4%	9.8%	8.5%	6.7%	7.7%
Secondary net enrollment rate (2015)	37.9%	25.4%	25.9%	35.5%	38.5%
Secondary repetition rate (2014)	4.9%	3.7%	4.9%	4.2%	4.3%
Secondary drop-out rate (2014)	8.5%	12.4%	11.5%	7.1%	8.0%

Source: Educational statistics of the Ministry of Education in El Salvador.

In the four departments, there were 1,344 basic education public schools, of which 606 basic education public schools had at least both cycle 1 (grades 1 to 3) and 3 (grades 7 to 9). Around 25 percent of public basic education schools in the four departments operate with only one or two teachers;

and allocated the necessary budget. Although technical cooperation from JICA was completed, the Ministry continued the activities.

[10] Though the sampling frame was composed of basic education public schools which had cycles 1 and 3, it was eventually found that all the schools also had cycle 2.

these were not included in the sampling frame. The country suffers from security problems due to the presence of gang members inherited from past civil conflicts. Intentional homicides per 100,000 were 61.8 in 2017 (World Bank 2020), the highest in the world. The schools were also affected by the activities of gang members (USAID 2017). Schools located in areas severely affected by such activities, and any that were physically difficult to access were excluded from the sampling frame. Outside the experimental design, the Millennium Challenge Corporation (MCC) planned to distribute ESMATE textbooks in 2018. The schools receiving intervention from MCC were also excluded from the evaluation framework. As a result, the sampling frame was comprised of 369 basic education public schools. From the sampling frame, 250 basic education public schools were randomly sampled, half of which (125 schools) were randomly assigned to the treatment group while the other half were assigned to the control group (Table 2).[11] Stratification variables in the random sampling and treatment assignment of schools were their department and rural/urban designation. If there were several classes of the targeted grades in the sampled school, one class was randomly selected. For security reasons, survey teams conducted field surveys during the morning shift at all schools that offered one or in the afternoon if necessary.

Table 2: Sampling frame of schools in the four departments

	Cabañas	La Union	San Miguel	San Vicente	Total
(A) No. of public schools (primary and/or lower secondary)	265	375	468	236	1,344
(B) Schools with cycle 1 and cycle 3 in (A)	104	144	247	111	606
(C-1) Schools without difficulty in access or security in (B)	64	68	164	105	401
(C-2) Schools not targeted by the MCC program (Sampling frame)	64	49	151	105	369
(D-1) Sampled schools (Treatment)	22	16	51	36	125
(D-2) Sampled schools (Control)	21	17	51	36	125

Note: Data source is the baseline survey of this research, and the educational statistics of the Ministry of education in El Salvador.

[11] The sample size was calculated with the following conditions: minimum detectable effect size: 0.2 standard deviations of test scores; cluster size: 20 students on average; significance level: 0.05; power: 0.8; and intra-cluster correlation coefficient: 0.25. Considering the risk of attrition of schools because of security issues, 12 schools were added respectively to the treatment and the control groups. The actual value of the intra-cluster correlation coefficient of the end-line scores is around 0.30.

In the baseline survey, seven schools in the treatment group and four schools in the control group were excluded for security reasons (Koei Research & Consulting Inc. 2018).[12] In addition to these eleven excluded schools, there were no students enrolled in grade 2 at one school in the treatment group. Based on the educational census survey data from the Ministry of Education in El Salvador, the sampling frame is compared with the original population of schools in the four departments. The results are reported in Table A in Appendix 2 of this chapter.

2.2.3 Assessment of math learning level

To assess the math learning of students, all three rounds of surveys (baseline, end-line, and follow-up) conducted written tests. The baseline survey was conducted from January to March 2018, the end-line survey from September to October 2018, and the follow-up survey from September to October 2019. The school year in El Salvador starts in mid-January and finishes in mid-November. To account for the progress following the curriculum, test items differ across these three tests. The tests were designed to measure student learning of math content defined in the curriculum. The test administered during the baseline survey assessed math content learned in the 1st grade. The test given at the end-line survey assessed math content learned in the 2nd grade. Each of these tests consisted of 20 questions, including problems posed in texts. The test given with the follow-up survey assessed math content learned in the 3rd grade.

Each test assessed different cognitive skills (knowing, applying, and reasoning) and cognitive domains (number and operation, quantity and measurement, and geometry). The number of items per unit of math content was defined according to the volume of lesson hours necessary for the unit. The compositions of test items are presented in Tables B.1 to B.3 in Appendix 3-1. The duration of the math tests was 45 minutes. The tests were administered by survey teams without the presence of teachers in the classroom. The test of the follow-up survey consisted of 25 questions, five of which

[12] At the end-line survey, three schools in the control group were additionally excluded because of security reasons (Koei Research & Consulting Inc. 2019).

were the same as questions given at the end-line survey.[13] Based on the common five math items in the end-line and the follow-up surveys, the test scores in the two rounds are linked using the item response theory. The process of linking test scores is reported in Appendix 3-2 of this chapter.

Those 2nd grade students who repeated the same grade in the following school year were kept in the sample and given the math test for 3rd grade students in the follow-up survey. In the follow-up survey data, 1.68 percent of students in the treatment group and 1.3 percent of students in the control group repeated the 2nd grade.

2.2.4 Comparison of student, teacher, and school characteristics

The balance of student characteristics is checked by the following equation

$$Y_{ik0}=\alpha_0+\delta_0 Treatment_k + D_k\beta_{0D}+\varepsilon_{ik0} \quad (1),$$

where Y_{ik0} represents the student characteristics such as math test baseline score for student *i* in school *k*. Test scores are standardized by the mean and standard deviations of the scores among students belonging to the control group. *Treatment* is an assignment to the treatment group. D_k is a vector of strata dummies constructed by the stratification variables in the random assignment, i.e., department dummies and the rural / urban dummy of school *k*. Robust standard errors are clustered at the school level.

Characteristics of students, teachers, and schools in the treatment and control groups are presented in Tables 3-1 and 3-2. The difference in standardized test scores is 0.08 and not statistically significant (the p-value is 0.149). While the percentage of students who had math textbooks when they were in the 1st grade was slightly higher in the control group than the treatment group, the other characteristics are similar between the two groups. In terms of the characteristics of teachers and schools, a difference in teacher qualification that is statistically significant is observed, but systematic differences are not observed between the two groups. The tables show that the treatment and control groups are well balanced, indicating successful randomization.

Because of logistical reasons, it was not possible to conduct the baseline survey before any

[13] All the test items are different between the baseline and end-line tests. Common five items in the end-line and the follow-up surveys are the item No. 9, 10, 11, 13, and 19 (the end-line) and No. 1 through 5 (the follow-up).

component of the intervention package started. The baseline survey was started in mid-January 2018, just after the distribution of textbooks to students at the beginning of a school year.[14] Consequently, it would be plausible to think that the intervention might have already affected student math learning measured in the baseline because it was conducted just after the distribution of textbooks. The kernel density curves of the baseline scores (Figures 1-1 and 1-2) show that the students with the lowest quartile baseline scores in the treatment group obtained slightly better scores than those in the control group. There were two items (No. 2 and No. 4) that students with the lowest quartile baseline scores in the treatment group correctly answered more successfully. Thus, this study also uses an alternative definition of the baseline score that excludes these two items. The difference in baseline math test scores excluding those two items is 0.06 and not statistically significant (the p-value is 0.241) (Table 3-1).

Figure 1-1. Density curves of Z scores (20 test items) at the baseline

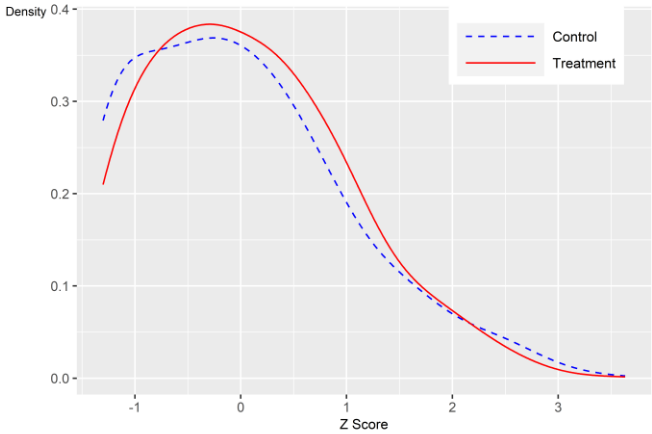
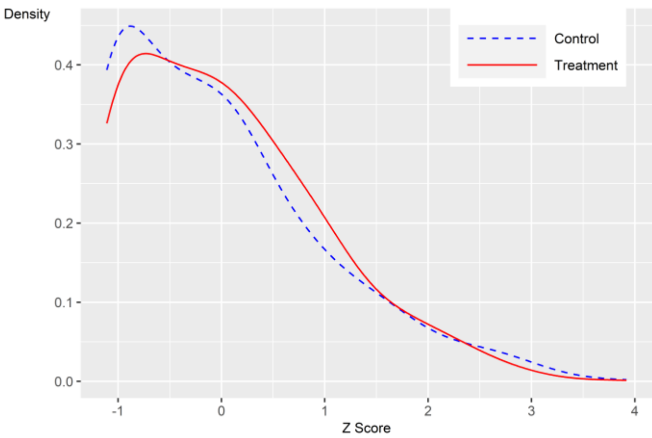


Figure 1-2. Density curves of Z scores (18 test items) at the baseline



Note: Data source is baseline survey of this research.

[14] The survey started on January 20 and finished on March 1, 2018.

Table 3-1. Comparison of student characteristics

Content	Treatment (1)	Control (2)	Diff. (3): (1)-(2)
Morning Shift (%)	94.94	91.22	3.72
Age	7.83	7.71	0.07
Standard deviation	(0.83)	(0.79)	
Sex (Male) (%)	49.61	51.62	-2.01
Number of elder brother or/and sister	1.62	1.57	0.04
Standard deviation	(1.83)	(1.79)	
Number of younger brother or/and sister	0.83	0.82	0.01
Standard deviation	(0.94)	(0.89)	
Math test score			
Raw test scores (Total point: 20)	5.05	4.75	0.30
Standard deviation	(3.41)	(3.65)	
Standardized test scores (20 items)	0.08	0.00	0.08
Raw test scores (Total point: 18, not including Q2&Q4)	3.72	3.53	0.20
Standard deviation	(3.00)	(3.18)	
Standardized test scores (18 items)	0.06	0.00	0.06
Asset of study			
Math textbook last academic year (%)	30.44	35.97	-5.53*
Math notebook last academic year (%)	87.36	87.81	-0.45
Own Study Desk at Home (%)	32.18	33.21	-1.03
Asset of student household			
Smartphone (%)	74.87	74.81	0.06
Computer (%)	24.77	22.48	2.29
Refrigerator (%)	82.77	80.88	1.89
Car (%)	31.79	30.61	1.18
TV (%)	90.92	91.87	-0.96
Tap water (%)	79.41	80.28	-0.87
Electricity (%)	95.15	94.42	0.73
Flush Toilet (%)	56.09	52.38	3.71
Use wood for cooking (%)	57.33	58.94	-1.61
Student household economic status index	0.036	-0.037	0.07
Number of schools	117	121	
Number of students	1,939	1,846	

Note: (1) Data source is the baseline survey of this research. Student household economic status index is the first component in the principal component analysis on different types of household assets. For more detail, please refer to Appendix 4 of this chapter. (2) * p<0.1. Column 3 reports the difference between the treatment and control groups and the result of the test for the difference in means between the two groups. The test controls for strata fixed effects constructed by the stratification variables in the random assignment (department and urban status). Robust standard errors are clustered at the school level.

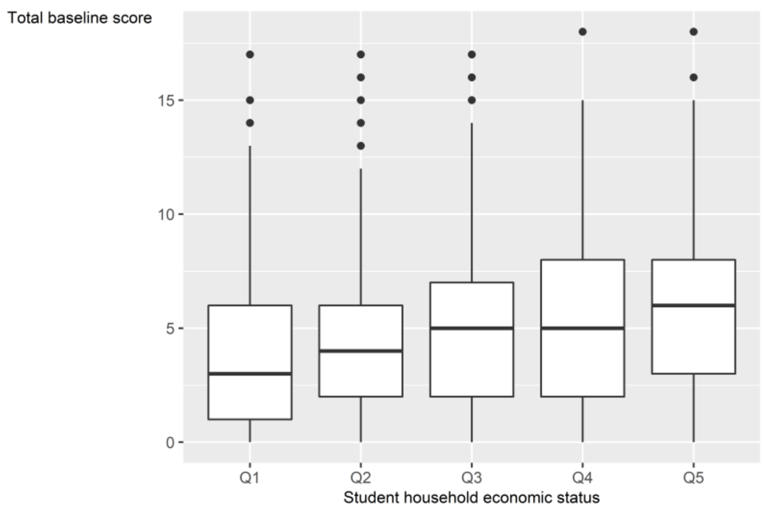
Table 3-2: Comparison of teacher and school characteristics

Content	Treatment (1)	Control (2)	Mean Diff. (3): (1)-(2)
Teacher characteristics			
Sex (Male: 1) (%)	22.2	24.8	-2.57
Age	45.9	47.1	-1.17
Standard deviation	(9.4)	(7.8)	
Total teaching period (years)	22.3	22.6	-0.33
Standard deviation	(8.4)	(8.2)	
Teaching other subjects in grade 2 (%)	88.9	90.1	1.19
Highest Academic Degree			
High school (%)	6.84	4.13	2.70
Professorate (%)	70.9	75.2	-4.27
Bachelor (%)	20.5	19.0	1.50
Master (%)	0.00	0.00	0.00
Doctor (%)	0.00	0.00	0.00
Teacher qualification (1)			
Pedagogical Bachelor (%)	13.7	6.61	7.06*
Professorate (%)	74.4	81.8	-7.46
Bachelor in Education (%)	14.5	9.91	4.61
Master in Education (%)	0.00	0.00	0.00
Doctorate in Education (%)	0.00	0.00	0.00
Pedagogical Training Course (%)	1.71	4.96	-3.25
Teacher qualification (2)			
Basic Education Teacher (cycle 1 and 2) (%)	47.0	51.2	-4.23
Mathematics Specialty Teacher (cycle 3) (%)	7.69	7.43	0.25
Teacher specialized in other than math (cycle 3) (%)	19.7	23.1	-3.48
Others (%)	29.1	24.8	4.27
School characteristics			
Number of students			
Number of students (2nd grade) in the morning shift	24.6	22.0	2.62
Standard deviation	(18.4)	(14.6)	
Number of students (2nd grade) in the afternoon shift	24.4	19.6	4.74
Standard deviation	(11.7)	(9.8)	
Number of students (Total)	258	235	24.2
Standard deviation	(232)	(177)	
Repetition and dropout rate (%)			
Repetition rate (morning shift of 2nd grade last academic year)	3.44	2.66	0.79
Repetition rate (afternoon shift of 2nd grade last academic year)	3.62	3.36	0.25
Dropout rate (morning shift of 2nd grade last academic year)	4.70	5.57	-0.09
Dropout rate (Afternoon shift of 2nd grade last academic year)	7.74	4.34	3.40
Number of teachers	10.7	10.1	0.61
Standard deviation	(9.2)	(6.8)	
School facility (%)			
Electricity	100	100	0.00
Drinking Water	80.3	86.0	-5.61
Computer	94.9	92.6	2.31
Library	25.6	24.8	0.09
Laboratory	12.0	6.61	5.35
Donor support within 5 years (except ESMATE) (%)	91.5	93.4	-1.94
Number of schools	117	121	

Note: (1) Data source is the baseline survey of this research. (2)* p<0.1. Column 3 reports the difference between the treatment and control groups, and the result of the test for the difference in means between the two groups. For binary values, the chi-squares test was conducted with stratified data (department and urban status). For the number of students, repetition rate, and dropout rate, the Wilcoxon rank sum test was conducted with stratified data (department and urban status).

Through student interviews, the baseline survey collected information on different types of student household assets. A composite index on student household economic status is computed from the data on different types of household assets by the principal component analysis. The first principal component accounts for around 25 percent of the overall variance of the variables on different types of household assets and all of their loading coefficients are positive. Therefore, the first principal component is taken as the index of student household economic status. The details of the principal component analysis are reported in Appendix 4. The whole sample is divided in quintiles according to the first principal component, and the correlation of math test scores and household economic status is checked. The boxplots in Figure 2 show that students with higher household economic status obtain higher math scores. The academic achievement of students is moderately correlated with their household economic status.

Figure 2. Boxplots of total baseline test scores by the student household economic status



Note: Student household economic status is represented by the first principal component of different types of student household assets. Students are grouped by 20 percent according to the first principal component. Q1 represents the bottom 20 percent, and Q5 includes the highest 20 percent.

2.2.5 Student attrition

The same students were tracked through three rounds of surveys in 2018 (baseline and end-line) and 2019 (follow-up). Of the 3,785 students present at the baseline survey, 753 were absent at the end-line survey, and 912 were absent at the follow-up survey. The attrition rate was 18.6 percent at the end-line survey and 23.3 percent at the follow-up survey in the treatment group. In the control group, the attrition rate was 21.3 percent at the end-line survey, and 24.9 percent at the follow-up survey.[14] Whether differential attrition occurred between the two groups is checked by regressing the student attrition dummy on the treatment assignment dummy, student characteristics, and strata dummies constructed by stratification variables (department, urban/rural designation). The results, shown in Table D-1 in Appendix 5, indicate that the attrition was not differential between the two groups.[15] The results also show that attrition tended to occur for students with lower academic achievement and higher age.

2.2.6 Estimation strategy of the impacts of the ESMATE program on student learning[16]

The impacts of the ESMATE program on student math learning in year 1 and year 2 are estimated by the following equation,

$$Y_{ikt}=\alpha_t+\gamma_t(Y_{ik0}\times Year_t)+\delta_t(Treatment_k\times Year_t)+C_{ikt}\beta_c+ P_{mkt}\beta_p+ S_{kt}\beta_s+ D_k\beta_D+\varepsilon_{ikt} \quad (2),$$

where Y_{ikt} on the left-hand side (t=1: end-line and t=2: follow-up) represents the math test score for student *i* in school *k* at the end-line or follow-up and Y_{ik0} on the right-hand side is his/her test score at the baseline. Baseline test scores (Y_{ik0}) are standardized by the mean and standard deviations of test scores of students belonging to the control group. Based on the common five math items in the end-line and the follow-up surveys, this chapter links the test scores on the left-hand side using the item response theory. The linked scores are called "IRT scores." To show the progress in math learning from year 1 to year 2, the IRT scores at the two rounds of survey are standardized by the mean and standard deviations of the IRT scores of the control group at the end-line survey. To evaluate the impact by the cognitive skills (i.e. knowing, applying, and reasoning), the sub-totals of test scores standardized

[14] There were students who participated in the baseline and follow-up surveys, but did not participate in the end-line survey. The number of students who participated in all three waves of the survey was 2,503.
[15] The balance of student characteristics is also checked in the samples who remained at the end-line and follow-up survey respectively, and confirmed the balance between the two groups. The results of the comparison are shown in Table D-2 and D-3 in Appendix 5.
[16] The pre-analysis plan of this research was registered at the following website on October 2018: https://www.socialscienceregistry.org/trials/3169. The estimation strategy was developed from the pre-analysis plan. The IRT scores were introduced to estimate the first-year impact and the accumulated impact.

by the mean and standard deviations of the control group at the round *t* of survey are used, instead of the IRT scores.

$Year_t$ is a vector of dummy variables of year *t* (t=1: end-line and t=2: follow-up). C_{ikt} is a vector of characteristics of student *i* at school *k* such as age, gender, the number of brothers and sisters, and characteristics of the family of student *i* such as different types of household assets at the baseline. C_{ikt} includes the shift at school (morning or afternoon) during year 1 and year 2, and whether student *i* repeated the 2nd grade in year 2. P_{mkt} is a vector of characteristics of teacher *m* at school *k,* who teaches mathematics to student *i,* such as age, gender, and educational qualification. S_{kt} is a vector of characteristics of school *k* such as school infrastructure and the total number of students attending school *k* at the baseline. S_{kt} includes characteristics of school principal at year 1 and 2. D_k is strata dummies constructed by the department and the urban/rural designation of school *k*. In equation (2), ε_{ikt} is the error term. Robust standard errors are clustered at the school level.

This study also investigates the heterogeneity of impact on student math learning outcomes by baseline scores and household economic status. Letting X_{ik0} stand for either the baseline score or the student household economic status, the estimated model becomes

$$Y_{ikt}=\alpha_t+\gamma_t(Y_{ik0}\times Year_t)+(\delta_{tA}+\delta_{tB}\,X_{ik0})(Treatment_k\times Year_t)+C_{ik}\beta_c+P_{mkt}\beta_p+S_{kt}\beta_s+D_k\beta_D+\varepsilon_{ikt} \quad (3).$$ [17]

If δ_{tB} in equation (3) is positive, the impact of the treatment is larger on students with a higher baseline score or higher household economic status. As shown in Figure 2, the baseline test scores and the student household economic status are positively correlated. Thus, equation (3) is expanded to include both interaction terms to estimate the heterogeneous impacts. The triple interaction terms of the treatment assignment, the baseline scores, and the household economic status are further added in the equation.

[17] When the heterogeneous impact by student household economic status is estimated, the interaction term of student household economic status and year dummy in the control variable are included in equation (3).

2.3. Results

2.3.1 The impacts of the ESMATE program on student math learning

Density curves of the IRT scores at the end-line survey show that the 2nd grade students in the treatment group as a whole improved math learning in year 1 (Figure 3-1). Since the IRT scores are comparable across the end-line and follow-up surveys, it is possible to track the progress of math learning in the two groups. Student math learning improved in both groups in year 2, but the students in the treatment group performed better (Figure 3-2). This indicates the accumulated impact of the first-year intervention in the following year.

Figure 3-1. Density curves of the IRT scores at the end-line survey

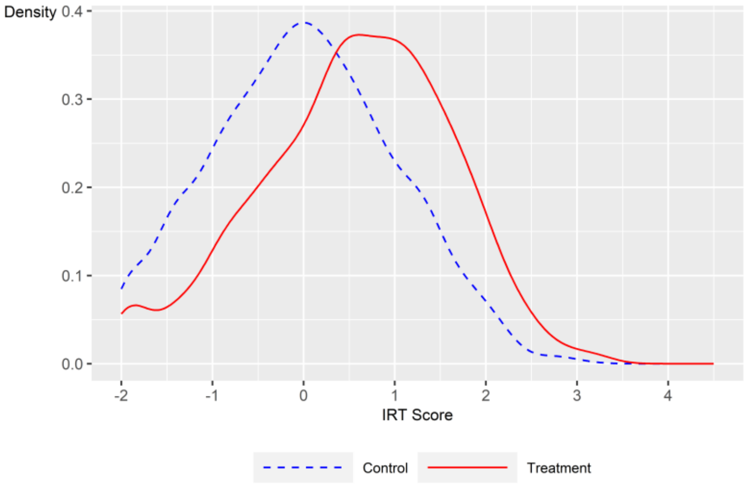
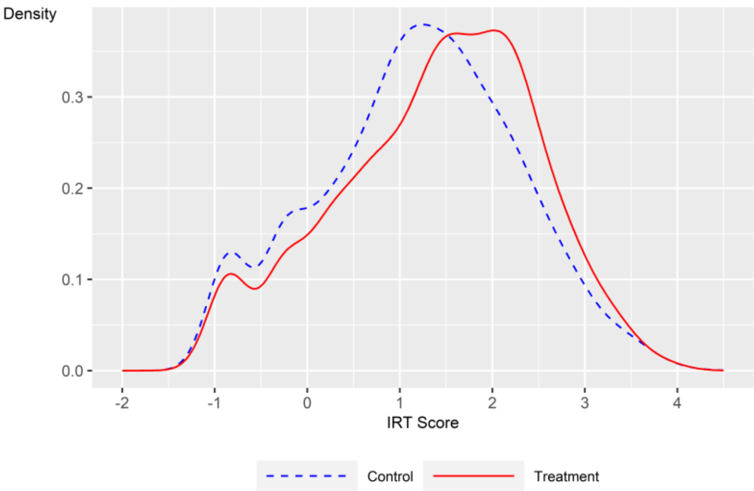


Figure 3-2. Density curves of the IRT scores at the follow-up survey



Note: Data source is the end-line and follow-up surveys of this research.
Number of students in Figure 3-1 is 3,032.
Number of students in Figure 3-2 is 2,878.

The average impacts of the ESMATE program on student math learning is estimated by equation (2), pooling the samples of students who took math tests both at the baseline and end-line surveys, and those who took math tests both at the baseline and follow-up surveys. The regression results are presented in Column (I)-1 to (I)-3 in Table 4-1. The average one-year impact of the ESMATE program is estimated at around 0.49 standard deviations of the IRT scores. The coefficient of the second-year dummy represents the improvement of math learning in the control group from year 1 to year 2. Then, the accumulated impact of the first-year intervention in the following year is estimated at around 0.13 standard deviations of the IRT scores.

The average impacts of the ESMATE program are estimated, using the samples of students who took math test at all three rounds of survey. The regression results using the panel data are shown in Column (II)-1 to (II)-3 in Table 4-1. The estimated impacts are slightly higher than the values estimated from the pooled data but remain at almost the same levels. The ESMATE program improved math learning of primary 2nd grade students in the treatment group; then the students advanced their math learning in the following year.

To analyze the impacts on student math learning from a different angle, the test items at the end-line and the follow-up surveys are divided respectively, according to the cognitive domains, i.e., knowing, applying, and reasoning. The sub-total of test scores is standardized by the mean and standard deviations of the control group at each round of surveys. Then, the one-year impact and the accumulated impact are estimated by the equation

$$Y_{ikt}=\alpha_t+\gamma_t Y_{ik0}+\delta_t Treatment_k+C_{ik}\beta_c+ P_{mkt}\beta_p+ S_{kt}\beta_s+ D_k\beta_D+\varepsilon_{ikt} \quad (4),$$

where subscript *t* represents the end-line (t=1) or the follow-up (t=2). Unlike equations (1)-(3), which are to be estimated by pooling both the end-line (t=1) and the follow-up (t=2) samples, equation (4) is estimated separately for the end-line and the follow-up. This is because without IRT, it is not possible to directly compare test scores in the end-line and the follow-up.

Table 4-1. The average impacts of the ESMATE program on student math learning

	(I)-1: Full sample	**(I)-2: Full sample**	**(I)-3: Full sample**	**(II)-1: Balanced panel sample**	**(II)-2: Balanced panel sample**	**(II)-3: Balanced panel sample**
Treatment×1st year dummy	0.575***	0.487***	0.494***	0.612***	0.519***	0.525***
	(0.077)	(0.072)	(0.071)	(0.077)	(0.072)	(0.071)
Treatment×2nd year dummy	0.219***	0.120*	0.128**	0.230***	0.132**	0.140**
	(0.079)	(0.062)	(0.061)	(0.080)	(0.063)	(0.062)
2nd year dummy	1.196***	1.233***	1.242***	1.222***	1.265***	1.273***
	(0.052)	(0.060)	(0.060)	(0.053)	(0.064)	(0.063)
Z baseline scores		0.467***			0.456***	
×1st year dummy		(0.035)			(0.036)	
Z baseline scores		0.535***			0.512***	
×2nd year dummy		(0.027)			(0.028)	
Z baseline scores (2)			0.452***			0.439***
×1st year dummy			(0.034)			(0.036)
Z baseline scores (2)			0.517***			0.492***
×2nd year dummy			(0.026)			(0.027)
Student characteristics	No	Yes	Yes	No	Yes	Yes
Teacher characteristics	No	Yes	Yes	No	Yes	Yes
School characteristics	No	Yes	Yes	No	Yes	Yes
Adj. R^2	0.235	0.449	0.444	0.245	0.451	0.445
N. obs.	5,909	5,909	5,909	5,006	5,006	5,006
N. Clusters	235	235	235	235	235	235

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

(1) Data source is three rounds of the survey of this research.

(2) Robust standard errors are clustered at the school level and are in parenthesis. Strata fixed effects are controlled in all the regressions but not shown.

(3) The IRT scores (end-line and follow-up) are standardized by the mean and standard deviations of the IRT scores of the control group at the end-line survey.

(4) Baseline test scores are standardized by the mean and standard deviations of the test scores of the control group. Z baseline scores (2) are the standardized test scores that exclude items No. 2 and No. 4.

The regression results are presented in Table 4-2 and 4-3. In year 1, regardless of the cognitive skills measured in assessment tests, the impact is positive and statistically significant, but the magnitude of the one-year impact of the ESMATE program is largest in the skill of knowing, followed by applying, and then reasoning.

While the accumulated impact of the ESMATE program in the following year is positive and statistically significant in the skill of knowing, the accumulated impacts on the skills of applying and reasoning are close to zero and not statistically significant. Students in the treatment group learned math content in the 3rd grade better than the control group did; but the accumulated gain did not lead to further advances in the skills of applying or reasoning.

Table 4-2. The one-year impact of the ESMATE program by the cognitive skills (year 1)

	(I)-1 Knowing	(I)-2 Knowing	(II)-1 Applying	(II)-2 Applying	(III)-1 Reasoning	(III)-2 Reasoning
Treatment	0.511^{***}	0.520^{***}	0.287^{***}	0.297^{***}	0.212^{***}	0.218^{***}
	(0.069)	(0.068)	(0.065)	(0.064)	(0.064)	(0.064)
Z baseline scores	0.430^{***}		0.487^{***}		0.316^{***}	
	(0.028)		(0.030)		(0.031)	
Z baseline scores (2)		0.416^{***}		0.487^{***}		0.314^{***}
		(0.027)		(0.029)		(0.031)
Num. obs.	3,030	3,030	3,030	3,030	3,030	3,030
N. Clusters	235	235	235	235	235	235

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

(1) Robust standard errors are clustered at the school level and are in parenthesis.
(2) Student test scores are standardized by the mean and standard deviations of control group test scores at each round of survey. Z baseline scores (2) are the standardized test scores that exclude questions No. 2 and No. 4.
(3) Student, teacher and school characteristics and strata fixed effect are controlled in all the regressions but not shown.

Table 4-3. The accumulated impact of the ESMATE programme by the cognitive skills (year 2)

	(I)-1 Knowing	(I)-2 Knowing	(II)-1 Applying	(II)-2 Applying	(III)-1 Reasoning	(III)-2 Reasoning
Treatment	0.131^{**}	0.141^{**}	0.055	0.063	0.019	0.025
	(0.062)	(0.061)	(0.058)	(0.059)	(0.050)	(0.050)
Z baseline scores	0.514^{***}		0.419^{***}		0.326^{***}	
	(0.029)		(0.029)		(0.022)	
Z baseline scores (2)		0.502^{***}		0.408^{***}		0.321^{***}
		(0.029)		(0.029)		(0.022)
Num. obs.	2,878	2,878	2,878	2,878	2,878	2,878
N. Clusters	235	235	235	235	235	235

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

(1) Robust standard errors are clustered at the school level and are in parenthesis.
(2) Student test scores are standardized by the mean and standard deviations of control group test scores at each round of survey. Z baseline scores (2) are the standardized test scores that exclude questions No. 2 and No. 4.
(3) Student, teacher and school characteristics and strata fixed effect are controlled in all the regressions but not shown.

2.3.2 The heterogeneity of impacts of the ESMATE program[18]

The heterogeneity of impacts by the baseline scores and the student household economic status is analyzed by equation (3). The results are shown in Table 5. In terms of the first-year impact, the coefficients of the interaction term with the baseline scores are positive, and the coefficients of the interaction term with the student household economic status are negative. But none of the estimated values of both coefficients are statistically significant (Columns (I) through (III) in Table 5). On the other hand, the coefficient of the triple interaction term of the treatment assignment, baseline score level, and household economic status is slightly negative and statistically significant at the 5 percent level, which indicates that the heterogeneous one-year impact by the baseline score level is larger for students with lower economic household status. In terms of the accumulated impact, all the coefficients of the interaction terms are close to zero and not statistically significant.

(3) The impacts of ESMATE program on teaching practices

The ESMATE program was designed to change teaching practices to improve student math learning. The textbooks were structured to facilitate teachers to organize math lessons in which students engage in learning mainly through working on math problems. Each page of the ESMATE textbook, which corresponds to a class, is structured along four steps: (a) present the theme of the lesson; (b) pose example problems; (c) explain the general principle; and (d) provide exercises. Through a set of interventions composed of introductory teacher training, lesson observation by the school principal, and periodical reviews of student math tests in the ESMATE program, teachers were expected to become conscious of the progress of student math learning. In addition to improving math teaching practices, the ESMATE program also tried to strengthen student's study at home through increasing the frequency of homework assignments. Since the workbooks were designed in

[18] In the baseline survey, average raw test scores were 5.16 points for female students and 4.65 points for male students. The average score was lower for male students than female students. Thus, in addition to the analysis of heterogeneous impact by baseline score and student household economic status, the heterogeneous impact by sex (for male students) is analyzed using equation (3). The heterogeneous impact is estimated at -0.013 (standard error: 0.063) for year 1 and 0.003 (standard error: 0.068) for year 2. Both of the two coefficients are not statistically significant.

correspondence with the ESMATE textbook, and the textbook includes exercise sections, teachers can utilize them to assign math homework regularly.

Table 5: Heterogeneity of the impacts of the ESMATE program on student math learning

	(I)-1	(I)-2	(II)	(III)
Treatment ×1st year dummy	0.495***	0.489***	0.489***	0.496***
	(0.073)	(0.072)	(0.072)	(0.071)
Treatment ×2nd year dummy	0.128**	0.129**	0.128**	0.128**
	(0.062)	(0.062)	(0.062)	(0.062)
Treatment ×1st year dummy	-0.027		-0.039	-0.036
× Household economic status	(0.030)		(0.031)	(0.031)
Treatment ×2nd year dummy	-0.024		-0.024	-0.024
× Household economic status	(0.027)		(0.027)	(0.027)
Treatment ×1st year dummy		0.089	0.099	0.106*
× Z baseline scores (2)		(0.062)	(0.063)	(0.063)
Treatment ×2nd year dummy		-0.006	0.000	0.001
× Z baseline scores (2)		(0.048)	(0.048)	(0.048)
Treatment ×1st year dummy				-0.037**
× Z baseline scores (2) × Household economic status				(0.019)
Treatment ×2nd year dummy				-0.005
× Z baseline scores (2) × Household economic status				(0.017)
2nd year dummy	1.251***	1.252***	1.252***	1.252***
	(0.060)	(0.060)	(0.060)	(0.061)
Z baseline scores (2)	0.454***	0.411***	0.405***	0.405***
×1st year dummy	(0.035)	(0.050)	(0.051)	(0.051)
Z baseline scores (2)	0.512***	0.514***	0.511***	0.511***
×2nd year dummy	(0.026)	(0.040)	(0.040)	(0.040)
Household economic status	0.031	0.018	0.038*	0.038*
×1st year dummy	(0.022)	(0.016)	(0.023)	(0.023)
Household economic status	0.057***	0.044***	0.057***	0.057***
×2nd year dummy	(0.019)	(0.015)	(0.019)	(0.019)
Adj. R^2	0.441	0.442	0.442	0.443
N. obs.	5,909	5,909	5,909	5,909
N. Clusters	235	235	235	235

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

(1) Robust standard errors are clustered at the school level and are in parenthesis.
(2) The IRT scores (end-line and follow-up) are standardized by the mean and standard deviations of the IRT scores of the control group at the end-line survey.
(3) Baseline test scores are standardized by the mean and standard deviations of the test scores of the control group. Z baseline scores (2) are the standardized test scores that exclude items No. 2 and No. 4.
(4) Student, teacher, and school characteristics and strata fixed effect (department and urban status) are controlled in all the regressions but not shown.
(5) Student household economic status is the first principal component of the different types of household assets.

This research conducted math lesson observations and interviews with teachers. The end-line survey conducted math lesson observations of 117 teachers in the treatment group, and 118 students in the control group. The same teachers were tracked at the follow-up survey. At the follow-up survey, math lesson observations were conducted for 94 teachers in the treatment group and 86 teachers in the control group. Among the 180 teachers, there were 83 in the treatment group in charge of grade 3 students and 66 in the control group in year 2 of this research.

Two surveyors visited each school. During the math lesson, one surveyor observed the teacher, and the other looked at students. The former took notes on the following seven points in the instructional routines of teachers: (a) teacher posed math problem; (b) teacher checked student notebooks; (c) teacher walked around the classroom to check notebooks; (d) teacher advised students to consult with each other; (e) teacher told students to check their answer; (f) teacher instructed students with the wrong answer to try again; and (g) teacher assigned homework. The other surveyor counted the length of time the students engaged in a lesson, i.e., the time in which at least half of the students in a class solved math problems, referred to their notebooks or textbooks, or consulted with each other (designated below as "student engaged time"). Surveyors collected information on the frequency of homework assignments through teacher interviews, and the length of time studying at home through student interviews. In year 1 of this research, while the percentage of students who had a math textbook (other than the ESMATE textbook) in the control group was around 20 percent, almost all the teachers in the group prepared math lesson plans regularly, referring to either existing math textbook, teacher's guidebook, etc.

With the data, the impacts of ESMATE program on teaching practices are estimated by the following equation,

$$Y_{mkt}=\alpha_t+\lambda_t(Treatment_k \times Year_t)+ P_{mkt}\rho_p+ S_{kt}\rho_s+ D_k\rho_D+\upsilon_{mkt} \quad (5)$$

where Y_{mkt} represents either the percentage of engaged time in a math lesson, the frequency of homework assignments, or the instructional routine of teacher *m* in school *k* at *t*th survey (end-line: 1; follow-up: 2). The value on the instructional routine is a dummy that takes the value 1 for the teacher who conducted each activity in the instructional routine raised in the previous paragraph. Y_{mkt} also take values on the frequency of lesson observations or suggestions that teacher *m* received from the school principal at school *k*. The other variables, *Treatment*, $Year_t$, P_{mkt}, S_k, and D_k are the same as in equation (2). The error term is υ_{mkt}.

The results of regression analysis on the percentage of student engaged time in a math lesson and the frequency of homework assignments are shown in Column (I) and (II) in Table 6. In year 1, the ESMATE program increased the percentage of student engaged time in a lesson by around 6 to 8 percentage points, the equivalent of 3 to 4 minutes in a 45-minute lesson (Column (I)). The frequency of homework assignments in a week was reported by teachers using four levels: (a) Never; (b) Once a week; (c) Two or three times a week; and (d) Four or more times a week. The ESMATE program increased the frequency of homework assignments by around 0.4 level points in year 1 (Column (II)). In the following year, the control group increased the percentage of student engaged time in a math lesson and the frequency of homework assignments after receiving the package of interventions. Then, the differences between the two groups disappeared in year 2 of this research.

Figure 4-1 shows the estimates of the average one-year impact of the ESMATE program on the instructional routine in math class. In year 1, the ESMATE program increased the percentage of teachers in the treatment group who walked around in a classroom to check student notebooks and assigned homework ((C) and (G) in Figure 4-1). On the other hand, in year 2, the difference in the percentage of teachers who walked around in the classroom becomes statistically insignificant ((C) in Figure 4-2). The accumulate impact on homework assignments in math lessons becomes negative in year 2 ((G) in Figure 4-2). The follow-up survey selected up to three grade 3 students randomly per school to check their workbooks (Koei Research & Consulting Inc. 2020). Taking an example in the first unit in math workbooks, sampled students in the treatment group solved at least one math problem on around 90 percent of the pages. But teachers checked around 30 percent of the pages that student worked on.[19] While students solved math items in the workbook, teachers did not check the work or provide feedback to students consistently.

[19] There are in total ten units in the student math workbook corresponding to the curriculum. The average percentage of pages that students worked on decreased as the units advanced. The same trend was observed in the average percentage of pages that teachers checked. While students had the answers in the workbook to check their own work, the average percentage of pages that students checked was less than the half of pages that they worked on.

Table 6. The impacts of the ESMATE program on the practices of teacher and school principal

	Percentage of engaged time in lesson		Frequency of homework assignment		Frequency of visit of school principal		Frequency of suggestion of school principal	
	(I)-1: Full sample	(I)-2: Balanced panel	(II)-1: Full sample	(II)-2: Balanced panel	(III)-1: Full sample	(III)-2: Balanced panel	(IV)-1: Full sample	(IV)-2: Balanced panel
Treatment×	7.025^{***}	8.369^{***}	0.438^{***}	0.400^{***}	0.024	-0.182	-0.070	-0.280
1st year dummy	(2.056)	(2.524)	(0.070)	(0.089)	(0.158)	(0.209)	(0.155)	(0.206)
Treatment×	-1.883	-3.015	0.014	0.045	-0.060	-0.147	-0.133	-0.261
2nd year dummy	(2.199)	(2.455)	(0.068)	(0.076)	(0.185)	(0.210)	(0.180)	(0.202)
2nd year dummy	6.265^{***}	8.133^{***}	0.444^{***}	0.362^{***}	-0.151	-0.236	0.145	0.112
	(2.011)	(2.322)	(0.076)	(0.092)	(0.168)	(0.202)	(0.170)	(0.210)
Average of the control group (year 1)	14.476	13.931	3.373	3.424	3.797	3.894	2.542	2.636
Adj. R^2	0.115	0.157	0.129	0.082	-0.017	-0.013	0.055	0.082
Num. obs.	415	298	415	298	415	298	415	298

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

(a) Teacher and school characteristics, and strata fixed effect (department and urban status) are controlled in all the regressions but not shown. Robust standard errors are used.
(b) Dependent variable in column (II) is the frequency of homework assignments in a week. The value of frequency of homework assignments in a week takes four levels: (1) Never; (2) Once a week; (3) Two or three times a week; and (4) Four or more times a week.
(c) The frequency of lesson observation by school principal takes five levels: (1) Never; (2) Once; (3) Twice; (4) More than three times during the year, but less than once a month; (5) More than once a month.
(d) The frequency of suggestions by the school principal takes four levels: (1) Never; (2) Once a year; (3) Once every semester; (4) Two or more times each semester.

Figure 4-1. The one-year impacts of the ESMATE program on the instructional routines

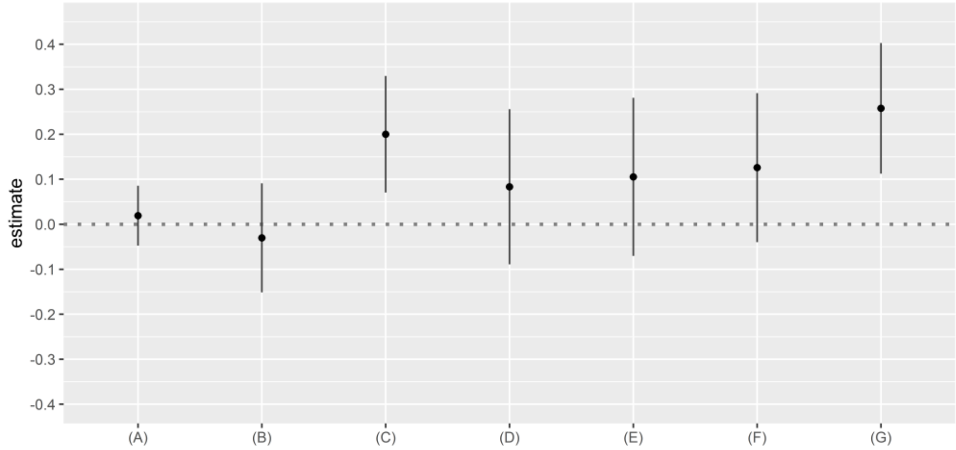
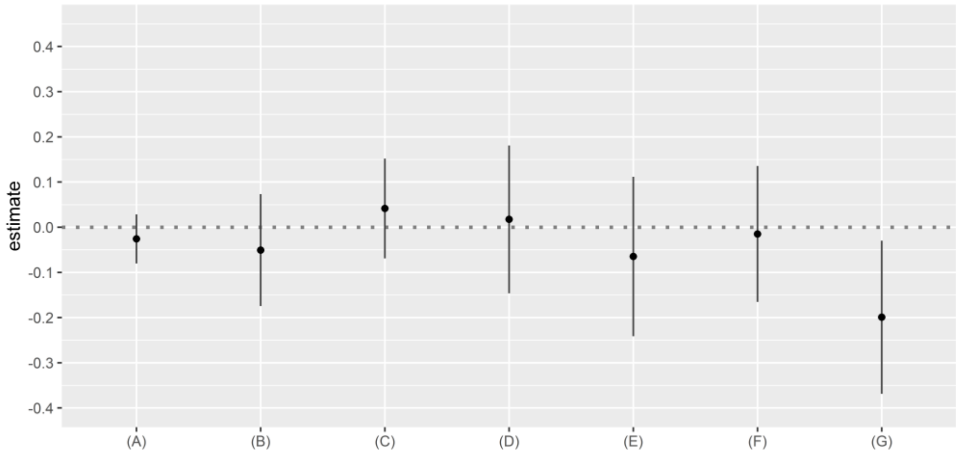


Figure 4-2. The accumulated impacts of the ESMATE program on the instructional routines in year 2



Note (1): The graph shows the point estimate and the confidence interval (95%) of the impact on instructional routine. The instructional routines are (A) Give exercise; (B) Check notebooks; (C) Walk in class to check notebooks; (D) Advise students to consult with each other; (E) Tell students to check their own answers; (F) Instruct students to correct wrong answers; and (G) Assign homework.
(2) The panel data of teachers who were in charge of grade 2 in year 1 and grade 3 in year 2 were used to estimate the impacts on the instructional routines. The total number of observations in the panel data is 298.

The ESMATE program tried to strengthen the support for teachers from the school principal through lesson observations. But the regression results demonstrate that the ESMATE program did not change the practices of school principals in terms of the frequency of lesson observations and suggestions for teachers (Column (III) and (IV) in Table 6).

2.4. Discussions

2.4.1 Differences in math achievement level among students in the treatment group after one-year intervention

While the ESMATE program improved student math learning on average, the differences in math achievement level among students in the treatment group became tangible. For example, after the one-year intervention, 27.1 percent of the 2nd grade students in the treatment group could not correctly respond to the two-digit addition item (“35+21”). Around half of the students in the group correctly answered the two-digit addition item above but could not solve the three-digit addition item (“253+174”). Then, the remaining students correctly responded to the three-digit addition item, but around a third of them could not solve the problem posed in texts that involved two-digit addition with carrying. The gap between curricular and learning levels would then become larger in upper grades.

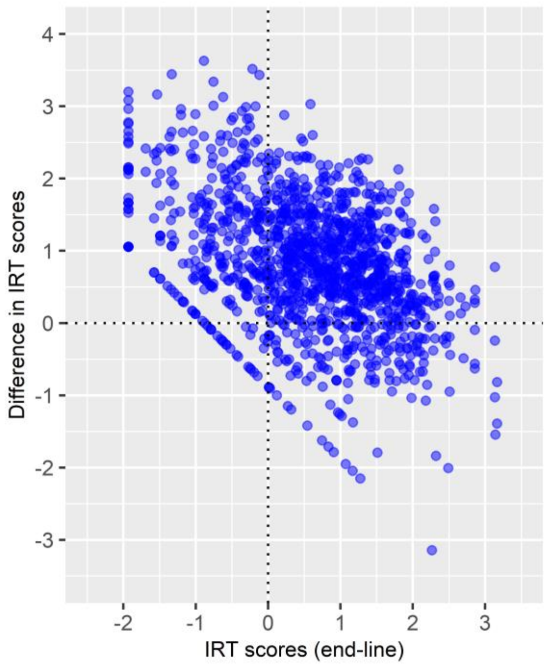
The ESMATE textbook is designed to increase the amount of time students engaged in math lessons and study at home, but the differences in math learning described above suggest that additional intervention strategies are required to address the different levels of math learning. For example, short-term remedial activities to support weak students would be an option to explore (Banerjee et al. 2007).

2.4.2 The correlates with the decrease in the IRT scores from year 1 to year 2

As shown in the section 2.3.1, the ESMATE program had an accumulated impact for students in the treatment group in year 2. Since the IRT scores are comparable across the end-line and the follow-up surveys, it is possible to check the trends in the progress of math learning from year 1 to year 2. Figure 5 is the scatter plot of the IRT scores of students in the treatment group at the end-line survey (X axis) and the IRT point difference between the end-line and follow-up surveys (Y axis). While a large percentage of students in the treatment group improved math learning from the end of 2nd grade to the end of 3rd grade, around 16 percent of the students decreased the IRT scores in the period.

To find the correlates with the decrease of the IRT scores in the treatment group, the difference in the IRT scores of student *i* at the two rounds of surveys is regressed on the student, teacher, and school characteristics, taking the sub-sample of the treatment group. One of the correlates with the decrease of the IRT scores in year 2 is the work assignment of teachers. In year 2, around 40 percent of teachers in the group were in charge of teaching math for several grades, because of multi-grade classes or/and double-shifts. The work assignment is negatively correlated with the difference in the IRT scores.[20] It might be difficult for those teachers to take sufficient time to support the math learning of sampled students in grade 3. The causes for the change in math achievement from year 1 to 2 is worth further research.

Figure 5. Scatter plots of the IRT scores at the end-line and follow-up surveys (Treatment group)



2.4.3 The cost-effectiveness of the ESMATE program

To understand the policy implications of this study, a cost-effective analysis of the ESMATE program is conducted, following the methodology presented by J-PAL (Bhula, R. et al. 2020; Dhaliwal et al. 2014). The cost-effectiveness is measured as the ratio of the aggregated impact of the

[20] The coefficient of the work assignment is -0.402 standard deviations of the IRT scores, which is statistically significant at the 1 percent level. Robust standard errors are clustered at the school level.

project (the average impact on student learning per student multiplied by the number of students impacted) to the aggregated cost of implementing the project. The cost-effectiveness is presented as the total standard deviations gained across the sample per 100 USD spent.[21] The cost-effectiveness of the ESMATE program in year 1, the total standard deviations gained across the sample per 100 USD spent, is estimated to be 4.06. The level of cost-effectiveness of the ESMATE program is comparable to or higher than the other similar programs cited in Kremer et al. (2013). When the accumulated impact in year 2 is taken into account, the level of cost-effectiveness will become higher.

2.5 Conclusions

School enrollment has rapidly increased since 1990 in developing countries at the primary level, but the quality of education has stagnated over the years. Improving learning is a critical issue for human and economic development in developing countries. The study by Hanushek and Woessman (2012) revealed that test scores that were higher by one standard deviation were associated with an average annual growth rate in GDP per capita that was about two percentage points higher over 40 years.

Textbooks can play key roles in improving student math learning. For example, textbooks are used to make annual teaching plans by teachers, and they describe the content and methodology of teaching and learning. Textbooks are also used for study at home by students. In developing countries, the availability of textbooks is limited, but simply distributing existing textbooks does not improve learning because of the mismatch between the levels of the curricula and student learning (Glewwe et al. 2009). This chapter investigated the impact of the package of interventions including the distribution of math textbooks that were carefully designed to improve math learning in El Salvador. The textbook in the package carefully subdivided subject contents, and the contents are sequenced to assure small-step learning by students.

While most impact evaluations focus on the impact on learning right after the intervention,

[21] Only the direct cost such as printing of teaching and learning materials and transportation fee for training is considered and the salary of the staff of the Ministry of Education is not taken into account. While travel allowances were not provided by the Ministry for teachers at the introductory training, the cost was included for the comparison of cost-effectiveness with other cases of intervention.

same students were tracked for two years to investigate the accumulated impact of the intervention in the following year. In this study, the math test scores in primary 2nd grade and 3rd grade were linked to capture the progress in learning from year 1 to year 2 of this research, using the item response theory. The ESMATE program improved student math learning by 0.49 standard deviations of the IRT scores in year 1. In the following year, students further advanced math learning by 0.13 standard deviations of the IRT scores. While the ESMATE program improved student math learning, the results revealed differences in achievement levels among students. Some students do not achieve the level supposed in the curriculum, and then have difficulty in learning math at the next grade. It is essential to further strengthen the program to help students continue their progress in math learning in primary education.

Appendix 1. Sample pages of teaching and learning materials developed by the ESMATE project

Appendix 2. Comparison of characteristics of schools

Appendix 3-1. Composition of math item

Appendix 3-2. Linking the test scores at the end-line and follow-up surveys

Appendix 4. Principal component analysis on different types of household assets

Appendix 5. Attrition analysis

References

Bashir, Sajitha, Marlaine Lockheed, Elizabeth Ninan, and Jee-Peng Tan. 2018. *Facing Forward: Schooling for Learning in Africa.* Washington DC: World Bank.

Banerjee, Abihijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics* 122(3): 1235–1264.

Bhula, Radhika, Meghan Mahoney, and Kyle Murphy. 2020. "Conducting cost-effectiveness analysis (CEA)." https://www.povertyactionlab.org/resource/conducting-cost-effectiveness-analysis-cea.

Bruns, Barbara and Javier Luque. 2015. *Great Teachers: How to Raise Student Learning in Latin America and Caribbean*. Washington DC: World Bank.

Bruns, Barbara, Leandro Costa, and Nina Cunha. 2018. "Through the looking glass: Can classroom observation and coaching improve teacher performance in Brazil?" *Economics of Education*

Review 64: 214–250.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. 2014. "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education." In *Education Policy in Developing Countries*, ed. Paul Glewwe. Chicago: University of Chicago Press.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children left behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1 (1): 112-35.

Glewwe, Paul (ed.) 2013. *Education Policy in Developing Countries*. Chicago: University of Chicago Press.

Hanushek, Eric A., and Ludger Woessmann. 2012. "Schooling, educational achievement, and the Latin American growth puzzle." *Journal of Development Economics* 99 (2): 497-512.

Jamison, Dean T., Barbara Searle, Klaus Galda, and Stephan P. Heyneman. 1981. "Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement." *Journal of Educational Psychology* 73 (4): 556–67

Japan International Cooperation Agency. 2019. *Project Completion Report of Project for the Improvement of Mathematics Teaching in Primary and Secondary Education in El Salvador. (in Japanese).*

Koei Research & Consulting Inc. 2018. *Impact Evaluation for the Project for the Improvement of Mathematics Teaching in Primary and Secondary Education in El Salvador: Baseline Survey Report*. JICA. Unpublished.

Koei Research & Consulting Inc. 2019. *Impact Evaluation for the Project for the Improvement of Mathematics Teaching in Primary and Secondary Education in El Salvador: End-line Survey Report*. JICA. Unpublished.

Koei Research & Consulting Inc. 2020. *Impact Evaluation for the Project for the Improvement of Mathematics Teaching in Primary and Secondary Education in El Salvador: Follow-up Survey Report*. JICA. Unpublished.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340: 297.

Kuecken, Maria and Marie-Anne Valfort. 2013. "When do textbooks matter for achievement? Evidence from African primary schools." *Economics Letters* 119: 311–15.

Marco, Gary L. 1977. "Item characteristics curve solutions to three intractable testing problems." *Journal of Educational Measurement* 14: 139-160.

McEwan, Patrik J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85 (3): 353– 94.

Ministerio de Educatión, El Salvador. 2014. EDUCACIÓN DE EL SALVADOR EN CIFRAS 2009-2014.

Ministerio de Educatión, El Salvador. 2016. BOLETÍN ESTADÍSTICO N° 19.

Ministerio de Educatión, El Salvador. 2016. Bases de Datos por Centros Escolares - Censo Escolar: Inicial Directorio 2016. El Salvador. https://www.mined.gob.sv/estadisticas-educativas/item/6116-bases-de-centros.html.

Ministerio de Educatión, El Salvador. 2018. Bases de Datos por Centros Escolares - Censo Escolar: Inicial Directorio 2018. El Salvador. https://www.mined.gob.sv/estadisticas-educativas/item/6116-bases-de-centros.html.

Ministerio de Educatión, El Salvador. 2018. Bases de Datos por Estudiante - Censo Escolar Inicial: Año 2018. https://www.mined.gob.sv/estadisticas-educativas/item/7154-bases-de-datos-spss.html.

Mullis, Ina V.S., Michael O. Martin, and Pierre P. Foy. (with John F. Olson, Corinna Preuschoff, Ebru Erberber, Alka Arora, and Joseph Galia). 2008. *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.

Pritchett, Lant, and Amanda Beatty. 2015. "Slow Down, You Are Going Too Fast: Matching Curricula to Student Skill level." *International Journal of Educational Development* 40: 276-288.

Read, Tony. 2015. *Where Have All the Textbooks Gone?: Toward Sustainable Provision of Teaching and Learning Materials in Sub-Saharan Africa*. Washington DC: World Bank.

Sabarwal, Shwetlena, David K. Evans, and Anastasia Marshak. 2014. "The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone." *Policy Research Working Paper*, 7021. Washington DC: The World Bank.

UNESCO Institute for Statistics (UIS). 2019. *UIS. Stat.* Retrieved from data.uis.unesco.org/

UNESCO 2017. *More than One-half of Children and Adolescents Are Not Learning Worldwide.Fact Sheet No. 46* (UIS/FS/2017/ED/46).

USAID 2017. *EVALUACIÓN DEL SECTOR EDUCACIÓN DE EL SALVADOR USAID/EL SALVADOR*.

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt and Emmanuel Jimenez. 2016. *The Impact of education programmes on learning and school participation in low- and middle- income countries*. London: International Initiative for Impact Evaluation.

World Bank. 2001. *A Chance to Learn: Knowledge and Finance for Education in Sub-Saharan Africa.* Washington DC: World Bank.

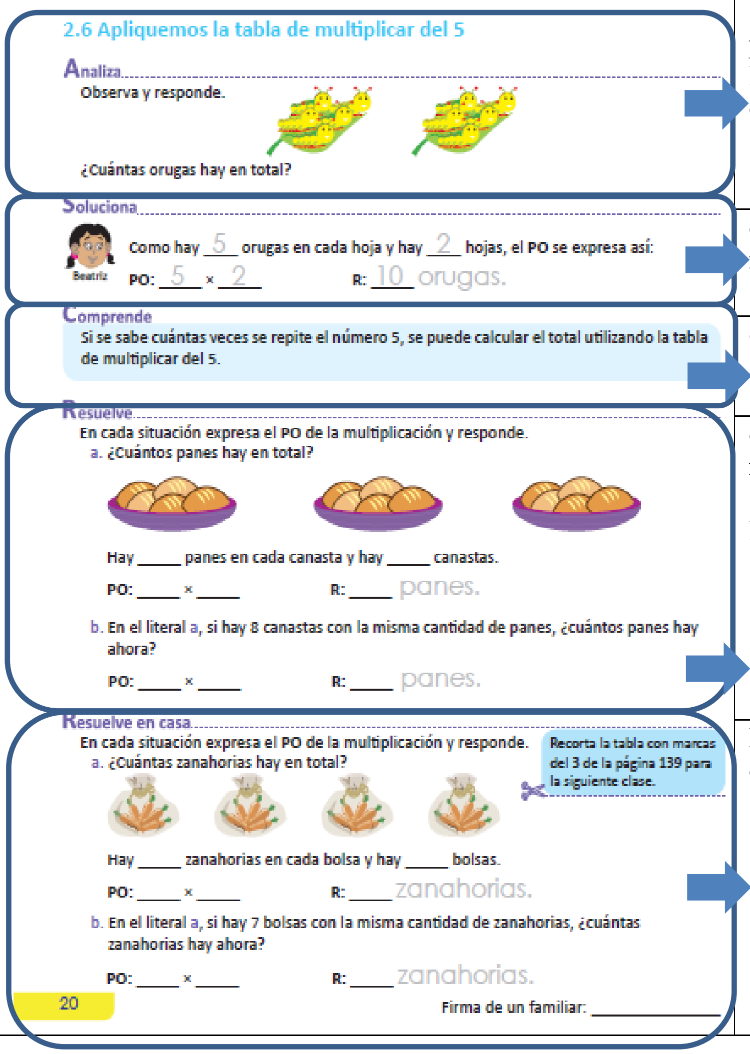
World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise.* Washington DC: World Bank.

World Bank. 2020. *World development indicators.* https://data.worldbank.org/indicator/

Sample pages of teaching and learning materials developed by the ESMATE project[i]

1. Textbook (primary 2nd grade)

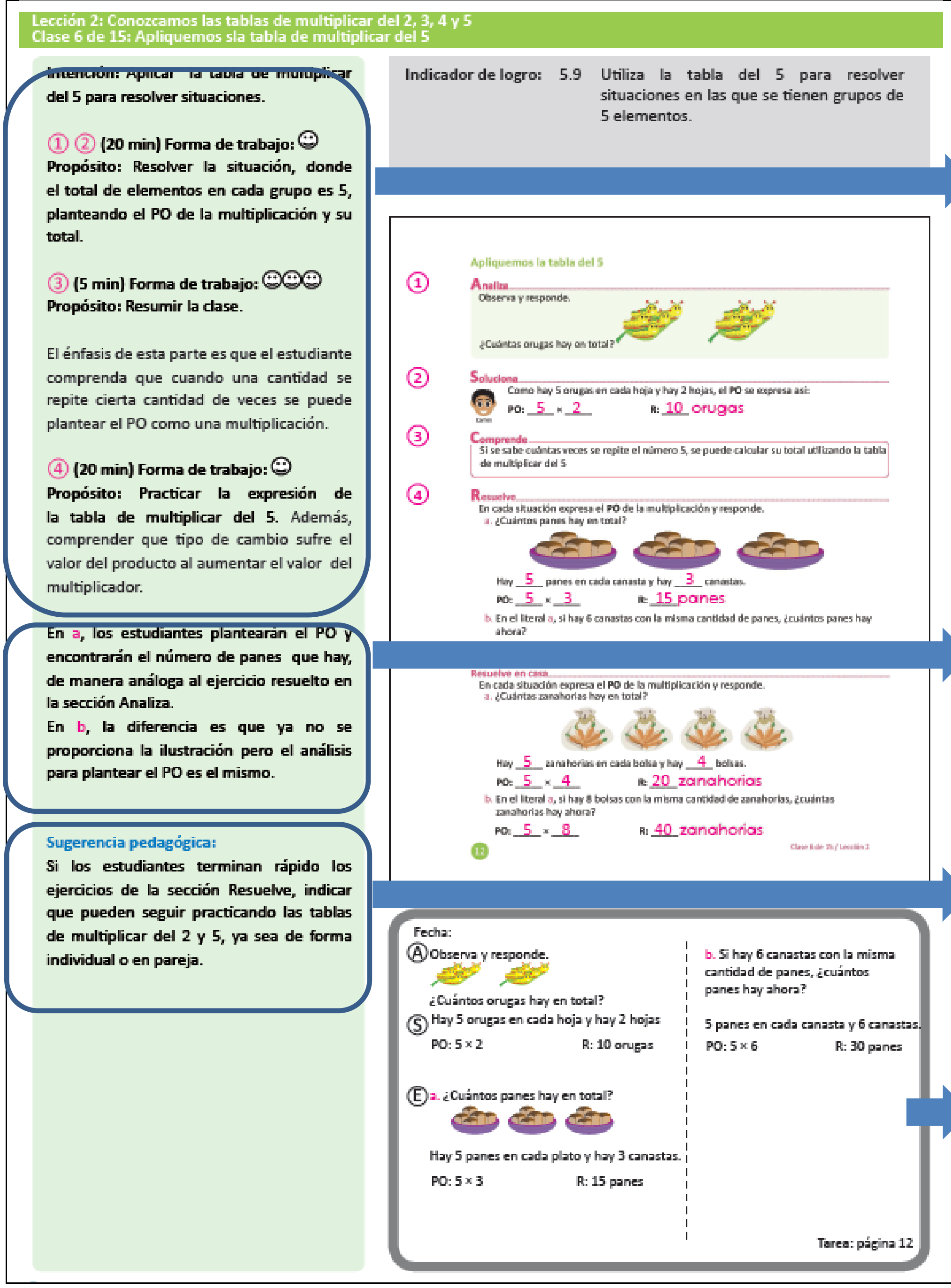
The standard learning process structured in the ESMATE textbook is (1) students understand the theme of the lesson with a concrete problem example, (2) they work on the problem example with the help of the teacher and explanation in the textbook, (3) they understand the mechanism (general principle) behind the example written in the textbook, and (4) they work on exercises in the textbook with support from the teacher.

Learning process	Sample page	By each part of textbook, teacher	By each part of textbook, student
(1) (2)	 2.6 Apliquemos la tabla de multiplicar del 5 Analiza Observa y responde. ¿Cuántas orugas hay en total?	poses the theme of lesson topic with concrete example	understands the theme of lesson topic
(3)	Soluciona Beatriz Como hay 5 orugas en cada hoja y hay 2 hojas, el PO se expresa así: PO: 5 × 2 R: 10 orugas.	checks student notebooks	solves example problem
	Comprende Si se sabe cuántas veces se repite el número 5, se puede calcular el total utilizando la tabla de multiplicar del 5.	explains how to solve	understands how to solve
(4)	Resuelve En cada situación expresa el PO de la multiplicación y responde. a. ¿Cuántos panes hay en total? Hay ____ panes en cada canasta y hay ____ canastas. PO: ____ × ____ R: ____ panes. b. En el literal a, si hay 8 canastas con la misma cantidad de panes, ¿cuántos panes hay ahora? PO: ____ × ____ R: ____ panes.	checks student notebooks and supports student learning	practices through solving problem
	Resuelve en casa En cada situación expresa el PO de la multiplicación y responde. a. ¿Cuántas zanahorias hay en total? Recorta la tabla con marcas del 3 de la página 139 para la siguiente clase. Hay ____ zanahorias en cada bolsa y hay ____ bolsas. PO: ____ × ____ R: ____ zanahorias. b. En el literal a, si hay 7 bolsas con la misma cantidad de zanahorias, ¿cuántas zanahorias hay ahora? PO: ____ × ____ R: ____ zanahorias. 20 Firma de un familiar: ____________	Instructs and checks homework	practices through solving problems at home

[i] Page 31 of JICA (2019) is translated in English and adapted by the author for this appendix.

2. Teacher's guidebook (primary 2nd grade)

Referring to the corresponding sample page of the textbook, technical guidance for teachers are provided in the guidebook. Teachers read the guidebook for their lesson preparation.

Sample page	Guidance in the guidebook for teachers
	- This part explains the purpose and time allocation of lesson - This part explains the intention of exercises. - This part suggests additional exercises for students - This part shows board plan

Lección 2: Conozcamos las tablas de multiplicar del 2, 3, 4 y 5
Clase 6 de 15: Apliquemos sla tabla de multiplicar del 5

Intención: Aplicar la tabla de multiplicar del 5 para resolver situaciones.

Indicador de logro: 5.9 Utiliza la tabla del 5 para resolver situaciones en las que se tienen grupos de 5 elementos.

① ② (20 min) Forma de trabajo:
Propósito: Resolver la situación, donde el total de elementos en cada grupo es 5, planteando el PO de la multiplicación y su total.

③ (5 min) Forma de trabajo:
Propósito: Resumir la clase.

El énfasis de esta parte es que el estudiante comprenda que cuando una cantidad se repite cierta cantidad de veces se puede plantear el PO como una multiplicación.

④ (20 min) Forma de trabajo:
Propósito: Practicar la expresión de la tabla de multiplicar del 5. Además, comprender que tipo de cambio sufre el valor del producto al aumentar el valor del multiplicador.

En a, los estudiantes plantearán el PO y encontrarán el número de panes que hay, de manera análoga al ejercicio resuelto en la sección Analiza.
En b, la diferencia es que ya no se proporciona la ilustración pero el análisis para plantear el PO es el mismo.

Sugerencia pedagógica:
Si los estudiantes terminan rápido los ejercicios de la sección Resuelve, indicar que pueden seguir practicando las tablas de multiplicar del 2 y 5, ya sea de forma individual o en pareja.

Apliquemos la tabla del 5

① Analiza
Observa y responde.
¿Cuántas orugas hay en total?

② Soluciona
Como hay 5 orugas en cada hoja y hay 2 hojas, el PO se expresa así:
PO: 5 × 2 R: 10 orugas

③ Comprende
Si se sabe cuántas veces se repite el número 5, se puede calcular su total utilizando la tabla de multiplicar del 5

④ Resuelve
En cada situación expresa el PO de la multiplicación y responde.
a. ¿Cuántos panes hay en total?
Hay 5 panes en cada canasta y hay 3 canastas.
PO: 5 × 3 R: 15 panes
b. En el literal a, si hay 6 canastas con la misma cantidad de panes, ¿cuántos panes hay ahora?

Resuelve en casa
En cada situación expresa el PO de la multiplicación y responde.
a. ¿Cuántas zanahorias hay en total?
Hay 5 zanahorias en cada bolsa y hay 4 bolsas.
PO: 5 × 4 R: 20 zanahorias
b. En el literal a, si hay 8 bolsas con la misma cantidad de zanahorias, ¿cuántas zanahorias hay ahora?
PO: 5 × 8 R: 40 zanahorias

Fecha:
Ⓐ Observa y responde.
¿Cuántas orugas hay en total?
Ⓢ Hay 5 orugas en cada hoja y hay 2 hojas
PO: 5 × 2 R: 10 orugas
Ⓔ a. ¿Cuántos panes hay en total?
Hay 5 panes en cada plato y hay 3 canastas.
PO: 5 × 3 R: 15 panes
b. Si hay 6 canastas con la misma cantidad de panes, ¿cuántos panes hay ahora?
5 panes en cada canasta y 6 canastas.
PO: 5 × 6 R: 30 panes
Tarea: página 12

Table A. Comparison of characteristics of schools

Content	Public schools in four departments (A)	Public schools that have cycle 1 and 3 (B)	Sampling frame (C)	Originally selected sample (D)	Surveyed schools (E)	P-value (D)=(E)
Percentage of N. of schools in urban area (2018)	18.6	25.0	29.4	29.2	30.7	0.72
Average N. of students (grade 2) (both shifts) (2018)	14.2	24.0	25.1	26.0	26.4	0.87
Average N. of total students (grade 1 to 9) (both shifts) (2018)	116.3	216.4	227.7	233.7	238.2	0.82
Percentage of Male students in grade 2 (2018)	51.9	52.0	51.7	51.4	51.4	0.99
Percentage of grade 2 students in morning shift (2018)	82.7	80.0	79.6	80.5	79.8	0.81
Percentage of grade 2 students repeated (2018)	3.5	3.2	3.0	3.2	3.2	0.99
Average age of grade 2 students (2018)	8.9	9.0	9.2	8.6	8.6	0.93
School infrastructure: Electricity (2016) (%)	97.0	99.0	98.9	98.8	98.7	0.95
School infrastructure: Water (2016) (%)	74.2	82.5	82.9	83.6	84.5	0.8
School infrastructure: Computer (2016) (%)	58.9	77.2	80.7	80.0	80.2	0.94
School infrastructure: Internet (2016) (%)	23.7	37.4	40.0	41.2	42.0	0.86
School infrastructure: Library (2016) (%)	15.3	23.0	24.2	24.0	24.8	0.84
School infrastructure: Laboratory (2016) (%)	6.1	8.6	9.2	9.6	9.7	0.98
N. of schools (of which census survey data is available)	1342	605	368	250	238	
N. of schools (which have grade 2 in 2018, and whose census survey data is available)	1226	601	364	249	238	

Note: (1) Data source is educational census survey data in El Salvador. Because the data of some schools are not available in the census survey data, the numbers of schools in Columns (A) to (C) in this table do not exactly match with those in Table 2 in Chapter 2. (2) p<0.1: *, p<0.05: **, p<0.01: ***. (3) Values on school facilities are binary (Yes:1, No:0). (4) The p-values on the number of students, percentage of students in morning shift, and percentage of students who repeated 2nd grade are the results of the Wilcoxon rank sum test with stratified data (department dummy, urban/rural dummy). (5) The p-values on binary values show the results of the chi-squares test with stratified data (department dummy, urban/rural dummy).

Table B.1. Composition of math item (Baseline)

Item No.	Item content	Cognitive domain	Cognitive skill
1	To write the number of quantity represented by the image	NO	Knowing
2	To add; one-digit number plus one-digit number = one-digit number	NO	Knowing
3	To subtract; one-digit number minus one-digit number = one-digit number	NO	Knowing
4	To find which side is longer and by how much using an eraser as unit of measure, in the notebook	QM	Knowing
5	To fill out the number pattern of 2 by 2	NO	Knowing
6	To choose the rectangle among different two dimensional figures	Geometry	Knowing
7	To choose the watch that indicates a determined time	QM	Knowing
8	To choose the container with more liquid through indirect comparison	QM	Knowing
9	To add; two-digit number plus two-digit number = two-digit number without carrying	NO	Knowing
10	To subtract; two-digit number minus two-digit number = two-digit number without borrowing	NO	Knowing
11	To fill out the number pattern of 6 by 6	NO	Knowing
12	To place numbers up to 50 on the number line	NO	Applying
13	To solve the written problem of addition; one-digit plus one-digit = two-digit number	NO	Applying
14	To solve the written problem of subtraction; two-digit minus one- digit = one-digit number	NO	Applying
15	To solve the written problem of multiplication; 7 times 5	NO	Applying
16	To solve the problem of equivalence between the amount of value and a few coins	QM	Applying
17	To find the unknown number in accumulation of the same number in an addition problem	NO	Applying
18	To solve the problem of accumulation of the same number adding one more time	NO	Reasoning
19	To solve the problem of which two operations are required and the result of the first operation is a number of the second operation	NO	Reasoning
20	To solve problem of addition in which one of the numbers is the ordinary number	NO	Reasoning

Note: NO: Number and Operation; QM: Quantity and Measurement

Table B.2. Composition of math item (End-line)

Item No.	Item content	Cognitive domain	Cognitive skill
1	To write three-digit number 109 represented by an image	NO	Knowing
2	To add with numbers up to two-digits, without carrying	NO	Knowing
3	To subtract numbers up to two-digit without borrowing	NO	Knowing
4	To measure an object in centimeters	QM	Knowing
5	To multiply 2 times one number	NO	Knowing
6	To choose the square among two dimensional figures	Geometry	Knowing
7	To identify the time of an event with determined duration	QM	Knowing
8	To read the capacity in liter	QM	Knowing
9	To add three-digit numbers with carrying to the position of hundreds	NO	Knowing
10	To subtract three-digit numbers with borrowing from the position of hundreds	NO	Knowing
11	To multiply 6 times one number	NO	Knowing
12	To place a three-digit number in the number line	NO	Applying
13	To solve the written problem of addition with the numbers up to two-digits with carrying to the position of tens	NO	Applying
14	To solve the written problem of the subtraction with the numbers up to two-digit with borrowing from the position of tens	NO	Applying
15	To write the multiplication from given addition of the accumulation with the same numbers and find the answer	NO	Applying
16	To solve the written problem of the situation of a purchase to find the combination of currencies to be given as the change	QM	Applying
17	To solve written problem of ratio	NO	Applying
18	To solve the written problem on the number to multiply	NO	Reasoning
19	To solve the problem of subtraction as an inverse operation of addition	NO	Reasoning
20	To solve an addition with three numbers, one of which is unknown	NO	Reasoning

Note: NO: Number and Operation; QM: Quantity and Measurement

Table B.3. Composition of math item (Follow-up)

Item No.	Item content	Cognitive domain	Cognitive skill
1	To add three-digit numbers with carrying to the position of hundreds	NO	Knowing
2	To subtract three-digit numbers with borrowing from the position of hundreds	NO	Knowing
3	To multiply 6 times one number	NO	Knowing
4	To solve the problem of the addition with the numbers up to two-digit with carrying to the position of tens	NO	Applying
5	To solve the addition with three numbers, one of which is unknown	NQR	Reasoning
6	To write four-digit number represented by an image	NO	Knowing
7	To add with four-digit numbers, with carrying one time	NO	Knowing
8	To subtract with four-digit numbers with borrowing one time	NO	Knowing
9	To measure an object in meters and centimeters	QM	Knowing
10	To multiply two-digit number and one-digit number without carrying	NO	Knowing
11	To choose the rectangle among the different two dimensional figures	Geometry	Knowing
12	To divide two-digit number by one-digit number without remainder	NO	Knowing
13	To represent the capacity writing the fraction	NO	Knowing
14	To multiply three-digit number and one-digit number with carrying one time	NO	Knowing
15	To divide two-digit number by one-digit number with remainder	NO	Knowing
16	To calculate the addition and subtraction combined with brackets	NO	Knowing
17	To read the number line of fractions	NO	Knowing
18	To interpret the bar chart	NQR	Knowing
19	To calculate the diameter from the radius of a circle	Geometry	Applying
20	To solve the written problem of the addition with four-digit numbers with carrying	NO	Applying
21	To solve the written problem of multiplication with three-digit numbers by one-digit number	NO	Applying
22	To solve the written problem of division with two-digit numbers by one-digit number	NO	Applying
23	To find the length of one edge from the total length of three edges of an equilateral triangle	Geometry / QM	Applying
24	To find a four-digit number, based on some hints using round up and off	NO	Reasoning
25	To find the correct answer of written problem of division in which it is required add 1 to the quotient	NO	Reasoning

Note: NO: Number and Operation; QM: Quantity and Measurement; NR: Numerical and Quantitative Relation

Appendix 3-2

Linking the test scores at the end-line and follow-up surveys

To estimate the values of item difficulty and item discrimination of the math items at the end-line and the follow-up surveys, the sample of students who were present at both surveys (N=2,503 students at each round of survey) was used. The two-parameter logistic model was employed to estimate the values.[ii]

The values of item difficulty and item discrimination of the math items at the end-line survey were estimated, the values at the follow-up survey ere estimated. As noted in section 2.2.3 in Chapter 2, there were five common math items in the two rounds of surveys. The five items were used for equating the values at the follow-up survey to those values at the end-line survey. For the equating, the Mean & Sigma method (Marco 1977) was used.

After the process of equating, the IRT scores of student *i* at the end-line survey, and at the follow-up survey were estimated respectively.

(1) Parameters of the math items at the end-line survey

No	Item difficulty	Item discrimination	No	Item difficulty	Item discrimination
1	0.036	0.948	11	-0.226	1.682
2	-0.489	1.612	12	0.791	1.782
3	0.076	1.525	13	0.757	2.012
4	-1.815	1.204	14	1.469	1.469
5	-0.831	1.712	15	0.569	1.719
6	0.042	1.049	16	0.649	1.550
7	0.193	1.055	17	0.858	1.875
8	0.162	0.624	18	3.705	0.976
9	1.025	1.857	19	1.153	1.488
10	1.704	1.521	20	3.299	1.072

[ii] R package called "IRTOYS" was used to estimate the parameters of the two-parameter logistic model, and the IRT scores.

(2) Parameters of the math items at the follow-up survey

No	Before equating		No	After equating	
	Item difficulty	Item discrimination		Item difficulty	Item discrimination
1	-0.386	1.529	1	0.636	1.754
2	1.052	1.290	2	1.890	1.480
3	-1.161	1.336	3	-0.040	1.533
4	-0.232	1.753	4	0.770	2.011
5	0.212	1.378	5	1.157	1.581
6	0.262	1.301	6	1.201	1.492
7	-0.069	1.665	7	0.912	1.910
8	1.327	1.366	8	2.129	1.567
9	1.407	0.978	9	2.199	1.122
10	0.330	1.705	10	1.260	1.956
11	0.498	0.950	11	1.407	1.089
12	0.142	1.645	12	1.097	1.887
13	2.387	1.029	13	3.053	1.181
14	0.927	2.187	14	1.780	2.509
15	0.988	1.840	15	1.833	2.110
16	1.684	1.607	16	2.440	1.843
17	0.363	1.109	17	1.289	1.272
18	0.716	1.497	18	1.597	1.717
19	1.606	1.450	19	2.372	1.663
20	0.949	1.637	20	1.800	1.878
21	1.803	1.804	21	2.544	2.069
22	1.553	1.580	22	2.326	1.812
23	2.260	1.501	23	2.943	1.722
24	3.705	1.413	24	4.202	1.621
25	4.158	0.990	25	4.596	1.136

Appendix 4

Principal component analysis on different types of household assets

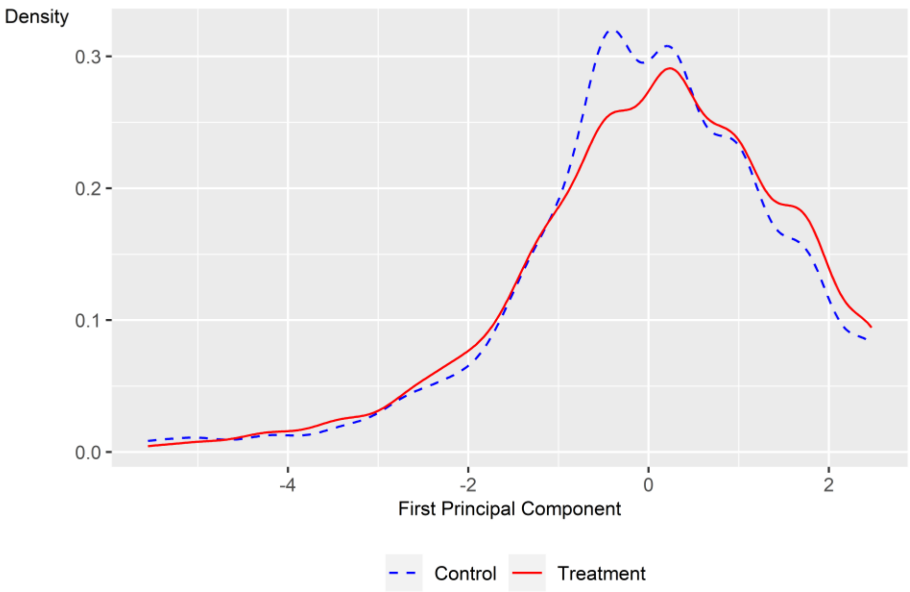
The baseline survey collected information on the following different types of student household assets through student interviews: (a) smartphone; (b) computer; (c) refrigerator; (d) car; (e) television; (f) access to tap water; and (g) flush toilet. In addition, the survey also collected information on access to electricity and the use of wood for cooking. The data on student household assets consisted of dummy variables that took the value of 1 when the student household possesses each category of asset that contributed to the wealth of a household. In the case of the cooking fuel, as the use of wood was an indicator of the lack of wealth (more wealthy households use gas or electricity for cooking), the non-use of wood in cooking was used as a dummy variable representing the asset. The descriptive statistics are presented in Table 3-2. Since the monetary value of household assets was not available, a composite index of student household economic status was computed by the principal component analysis, using the baseline survey data (N=3,785 students). The statistics of the first principal component are presented in Table C. The first principal component accounts for around 24.5 percent of the variances, and all of the nine asset types have positive loading coefficients, indicating that they contribute to the household wealth.

The kernel density curves are presented in Figure A. While a slight difference is observed in the density around the value 0 in the first principal component (x-axis), the difference between the two groups is not statistically significant, as reported in Table 3-1.

Table C. Statistics of first principal component

	Coefficient	Correlation with the first principal component
Smartphone	0.220	0.327
Computer	0.350	0.521
Refrigerator	0.396	0.588
Car	0.371	0.550
Television	0.366	0.543
Access to tap water	0.324	0.481
Electricity	0.343	0.508
Flush toilet	0.332	0.492
Non-use of wood in cooking	0.260	0.385

Figure A. Density curves of the first principal component



Attrition analysis

Whether differential attrition occurred between the treatment and the control groups was checked by regressing the student attrition dummy on the treatment assignment dummy, student characteristics, and strata dummies. The results are shown in Table D.1.

Table D.1. Sample attrition analysis

	End-line		Follow-up	
	OLS	**Logit**	**OLS**	**Logit**
(Intercept)	-0.049		-0.034	
	(0.109)		(0.095)	
Treatment	-0.024	-0.023	-0.016	-0.015
	(0.025)	(0.024)	(0.021)	(0.021)
La Union	-0.012	-0.013	0.042	0.046
	(0.038)	(0.045)	(0.043)	(0.048)
San Miguel	0.095**	0.100**	0.063	0.065
	(0.043)	(0.051)	(0.038)	(0.041)
San Vicente	0.055	0.063	0.027	0.029
	(0.039)	(0.049)	(0.033)	(0.037)
Urban	0.021	0.013	0.056	0.058
	(0.041)	(0.055)	(0.040)	(0.044)
La Union×Urban	0.033	0.046	-0.050	-0.049
	(0.057)	(0.079)	(0.064)	(0.056)
San Miguel×Urban	-0.012	-0.005	-0.033	-0.035
	(0.059)	(0.065)	(0.054)	(0.051)
San Vicente×Urban	0.008	0.016	-0.038	-0.038
	(0.061)	(0.074)	(0.052)	(0.051)
Z score baseline (2)	-0.039***	-0.042***	-0.040***	-0.043***
	(0.008)	(0.009)	(0.008)	(0.009)
Age	0.042***	0.036***	0.036***	0.033***
	(0.010)	(0.008)	(0.010)	(0.008)
Sex	0.018	0.018	0.006	0.006
	(0.013)	(0.013)	(0.013)	(0.013)
N. of elder brother/sister	0.005	0.005	0.005	0.005
	(0.004)	(0.004)	(0.004)	(0.004)
N. of younger brother/sister	-0.003	-0.004	0.004	0.004
	(0.006)	(0.006)	(0.007)	(0.007)
Total N. of different asset types	-0.010*	-0.009*	-0.005	-0.005
	(0.005)	(0.005)	(0.005)	(0.005)
Shift (Morning=1)	-0.112*	-0.116*	-0.042	-0.042
	(0.056)	(0.062)	(0.033)	(0.034)
Adi. R^2	0.033		0.016	
Num. obs.	3,783	3,783	3,783	3,783
F statistic	6.466		4.531	
Log Likelihood		-1818.62		-2043.57
Deviance		3637.25		4087.15

Table D.2. Characteristics of students (baseline) remained at the end-line survey

Content	Treatment (1)	Control (2)	Mean Diff. (3): (1)-(2)
Morning Shift (%)	96.32	91.40	4.92 **
Age	7.77	7.66	0.08
Sex (Male) (%)	48.35	50.89	-2.54
N. elder brother/sister	1.59	1.54	0.05
N. younger brother/sister	0.81	0.82	0.00
Test scores			
Raw test scores (Total points: 20)	5.22	4.97	0.25
Standard deviations	3.42	3.69	
Standardized test scores (Item: 20)	0.07	0.00	0.07
Raw test scores (Total points: 18, not including Q2&Q4)	3.87	3.71	0.15
Standard deviations	3.02	3.24	
Standardized test scores (Item: 18)	0.05	0.00	0.05
Asset of study			
Math textbook last academic year (%)	31.14	37.76	-6.62 **
Math notebook last academic year (%)	87.44	88.03	-0.59
Own Study Desk at Home (%)	32.19	33.70	-1.51
Asset of student household			
Smartphone (%)	74.89	74.42	0.47
Computer (%)	25.68	23.80	1.89
Refrigerator (%)	83.77	82.26	1.51
Car (%)	32.85	31.43	1.42
TV (%)	90.81	91.95	-1.15
Tap water (%)	79.33	81.02	-1.69
Electricity (%)	95.56	94.91	0.65
Flush Toilet (%)	56.25	52.68	3.56
Using wood for cooking (%)	57.32	58.60	-1.27
Student household economic status index	0.07	0.01	0.06
N. of schools	117	121	
N. of students	1,579	1,453	

Note: (1) Data sources are baseline and end-line surveys of this research. P<0.05: ** (2) Column 3 reports the difference between the treatment and control groups, and the result of the test for the difference in means between the two groups. The test controls for strata fixed effects constructed by the stratification variables in the random assignment (department and urban status). Robust standard errors are clustered at the school level.

Table D.3. Characteristics of students (baseline) remained at the follow-up survey

Content	Treatment (1)	Control (2)	Mean Diff. (3): (1)-(2)
Morning Shift (%)	107.39	111.03	-3.64
Age	7.79	7.73	0.06
Sex (Male) (%)	49.19	50.79	-1.60
N. elder brother/sister	1.60	1.53	0.06
N. younger brother/sister	0.81	0.81	0.00
Test scores			
Raw test score (Total points: 20)	5.27	4.93	0.34*
Standard deviations	3.43	3.66	
Standardized test scores (Item: 20)	0.09	0.00	0.09*
Raw test score (Total points: 18, not including Q2&Q4)	3.92	3.67	0.25
Standard deviations	3.04	3.21	
Standardized test scores (Item: 20)	0.08	0.00	0.08
Asset of study			
Math textbook last academic year (%)	31.07	37.18	-6.10*
Math notebook last academic year (%)	87.99	87.97	0.02
Own Study Desk at Home (%)	31.81	33.79	-1.98
Asset of student household			
Smartphone (%)	74.77	73.34	1.42
Computer (%)	25.37	23.13	2.24
Refrigerator (%)	84.36	82.20	2.16
Car (%)	33.09	31.27	1.82
TV (%)	91.01	91.50	-0.49
Tap water (%)	80.07	80.98	-0.91
Electricity (%)	95.44	94.60	0.84
Flush Toilet (%)	55.84	52.67	3.17
Using wood for cooking (%)	57.18	59.58	-2.40
Student household economic status index	0.08	-0.02	0.10
N. of schools	117	121	
N. of students	1,487	1,386	

Note: (1) Data sources are baseline and end-line surveys of this research. P<0.05: ** (2) Column 3 reports the difference between the treatment and control groups, and the result of the test for the difference in means between the two groups. The test controls for strata fixed effects constructed by the stratification variables in the random assignment (department and urban status). Robust standard errors are clustered at the school level.

Chapter 3. Strengthening Teacher Support for Students to Improve Math Learning: Empirical Evidence from a Structured Pedagogy Program in El Salvador

3.1 Introduction

A systematic review of evidence in education suggests that a structured pedagogy program that provides teaching and learning materials and other types of interventions such as teacher training is an effective approach to improving student learning (Snilsveit et al., 2016). To address several challenges regarding the improvement of learning (e.g., inadequately trained teachers and a lack of appropriate materials, curricula, and instructional approaches), a structured pedagogy program includes different types of components, such as the distribution of teaching and learning materials, teacher training, and instructional support from school principals (Snilsveit et al., 2016). However, given an extensive variety of possible combinations of interventions, systematic reviews have not clearly demonstrated the effective combination of different interventions for a structured pedagogy program, which would be an interest of policymakers (Piper et al., 2018). It is also necessary to consider possible complementarities among different inputs in a package of interventions (Kerwin and Thornton, 2020).

The experiment of this study in El Salvador intends to bridge the gap in understanding regarding the combination of different components in a structured pedagogy program at the lower secondary education level. While the Ministry of Education scaled up the ESMATE program nationwide for grades 7 through 9 (the lower secondary education level[22]), the Ministry could not cover several parts of the program, including the distribution of student math workbook and diagnostic math test, because of the budget constraints.[23] This study investigates the impact of the additional components in the

[22] In El Salvador, the primary (grade 1 through 6) and lower secondary education (grades 7 through 9) are compulsory.

[23] As introduced in Chapter 2, to improve student math achievement, the Ministry of Education implemented a project called "Project for the Improvement of Mathematics Teaching in Primary and Secondary Education" ("ESMATE project") with technical cooperation by Japan International Cooperation Agency (JICA). The ESMATE project developed a set of teaching and learning materials, i.e., teachers' guides, textbooks (ESMATE textbooks) and workbooks, then combined the provision of the materials with the other types of interventions as a structured pedagogy program ("ESMATE program").

ESMATE program on student math learning using a randomized controlled trial, and examine the mechanism by the causal mediation analysis.[24] The surveys tracked the same students for two years. During the first year, schools in the treatment group received the complete ESMATE program including the distribution of student math workbooks and diagnostic math tests, while a package of interventions that excluded these components was given to the control group. During the second year, there were no differences in the intervention between the two groups. Both groups received the ESMATE textbooks, teachers' guidebooks, and student math workbook.

Although the Ministry of Education scaled up the ESMATE program nationwide for grades 7 to 9 (the lower secondary education level),[25] it could not cover several interventions in the program due to budget constraints. This study focused on the interventions that were not covered in the nationwide scaling-up and investigates how the additional interventions could enhance the impact of the program on mathematics learning of grade 7 students using a randomized controlled trial.[26]

Additional interventions in the ESMATE program aimed to enhance the support of teachers for students to improve their mathematics learning in two ways. First, teachers in the treatment group developed annual math teaching plan at the introductory training. The Ministry also provided teachers with mathematics tests for students. In mutual review meetings during inter-semester periods, teachers would review and discuss the progress of annual teaching plan and student test results with colleagues. The process intended to make teachers conscious of the delay in the math curriculum and how much students learned math. The Ministry also provided initial on-site advice to school principals regarding mathematics lesson observation to increase the frequency of their lesson observations and suggestions (feedbacks) to teachers. The interventions aimed to improve teachers' teaching practices in their

[24] The randomized controlled trial was conducted under an agreement between the Ministry of Education in El Salvador and JICA on June and October 2018. The survey and database construction were done by Koei Research & Consulting Inc., under a contract with JICA, in collaboration with the Ministry of Education in El Salvador. Hitotsubashi University Research Ethics Examination Committee reviewed the research plan.

[25] In El Salvador, primary education (grades 1 through 6) and lower secondary education (grades 7 through 9) are compulsory.

[26] The randomized controlled trial was conducted with agreement between the Ministry of Education in El Salvador and JICA in June and October 2017. The survey and database construction were performed by Koei Research & Consulting Inc. under a contract with JICA, in collaboration with the Ministry of Education in El Salvador. Hitotsubashi University Research Ethics Examination Committee reviewed the research plan.

mathematics lessons. Second, in addition to the textbook, mathematics workbooks were distributed to students in the treatment group to strengthen their study at home. As the content of the workbook corresponded to the ESMATE textbook, teachers could use it for homework assignments.

This chapter examines the mechanism of the additional components in the structured pedagogy program through a review of the process data and a causal mediation analysis employing a method of Imai et al. (2010). The average causal mediation effect is the treatment effect that operates through a particular channel of a causal path. When there are several components in an intervention for the treatment group, it is hard to capture which component worked better through simple comparison with the control groups. But we can explore the mechanism using causal mediation analysis. In accordance with the additional components provided for the treatment group in year 1, the analysis takes three possible mediators: the provision of additional math class when the curriculum was delayed; students' time engaged in math learning in a lesson; and the frequency of homework assignments. Among the three mediators, only the average causal mediation effect of the provision of additional math class is positive and statistically significant.

This chapter also evaluates the accumulated impact of the first-year intervention on student math learning in the following year. Since students continue learning for years, it is critical to see whether students can advance their learning after receiving the intervention. This study tracked the same students for two years. To capture the progress in learning for two years, the item response theory (IRT) is used to link math test scores at the end of the 7th grade (lower secondary 1st grade) and the 8th grade. In 2019, both groups received the same interventions by the Ministry. The Ministry continuously provided a set of math textbooks, teacher's guidebooks, and student workbooks. In year 1, the additional components in the ESMATE program improved student math learning by 0.17 standard deviations of the IRT scores. Students in the treatment group responded correctly to the test items such as simple operation with positive and negative numbers, and direct and inverse proportions than the control group. The average accumulated impact in the following year was positive; however, it was not statistically significant. The gains in year 1 were not sufficient to yield an accumulated impact on math learning in year 2.

This chapter analyzes the heterogeneity of the impacts of the first-year impact with respect to student household economic status. In El Salvador, economic inequality is historically high. The Gini index was 54.5 in 1998 and gradually decreased with the modest economic growth of the country, reaching 38.6 in 2018 (World Bank, 2021). Survey data of this study confirms that student math baseline test scores positively correlate with the household economic status. Based upon the baseline data on different types of student household asset, a composite index that represents the student household economic status by the principal component analysis was constructed. Even after controlling the heterogeneous impact by the baseline score level, the impact of additional components in the ESMATE program was smaller for students with higher household economic status.

The remainder of this chapter is organized into the following five sections. Section 3.2 describes the experimentation design and the contents of the ESMATE program. Section 3.3 reports the impacts of the additional components in the ESMATE program on student math learning. Section 3.4 investigates the mechanism of the additional components by the causal mediation analysis, and report the results. Then, section 3.5 discusses the findings, and section 3.6 concludes.

3.2 Experimental design

3.2.1 Contents of the ESMATE program

In 2018, the Ministry of Education in El Salvador provided the schools in the control group with a package of interventions to improve students' mathematics learning. The primary component of the package was the distribution of the ESMATE textbooks and teachers' guidebooks. The ESMATE textbooks were designed to help students learn mathematics by solving mathematics problems. Each page included four lesson steps: (1) show the lesson topic, (2) pose example problems, (3) explain the general principle, and (4) provide exercises, so that teachers could organize daily mathematics lessons according to these four steps. The package also included other components, namely introductory training for teachers and school principals regarding how to use the textbook, and mutual review meetings among teachers to improve teaching with the ESMATE textbook. Introductory training for representatives of parent associations was also provided to strengthen family support for students' study at home.

For the treatment group, the Ministry provided supplementary interventions to the abovementioned package to strengthen the impact on students' mathematics learning in two ways. First, teachers in the treatment group developed annual math teaching plan at the beginning of school year in the introductory training.[27] The Ministry also provided teachers with mathematics tests for students. In mutual review meetings during inter-semester periods, teachers would review and discuss the progress of annual teaching plan and student test results with colleagues.[28] The Ministry also provided initial on-site advice for school principals regarding mathematics lesson observation to increase the frequency of their lesson observations and suggestions for teachers. The interventions aimed to improve teachers' teaching practices in mathematics lessons. Second, in addition to the textbook, mathematics workbooks that corresponded to the ESMATE textbook were also distributed to students in the treatment group. Since teachers could use the workbooks to assign homework regularly, the distribution of the workbooks would increase the frequency of homework assignments.

At the beginning of the school year in January 2018, the ESMATE textbooks and teachers' guidebooks were distributed to the treatment group. The distribution for the control group was delayed by approximately one month due to a delay in the Ministry's public procurement process. In the control group, the schedule for introductory training for teachers, school principals, and representatives of parent associations was also postponed for approximately one month, corresponding with the delayed

[27] Teachers in the control group also developed annual math teaching plan; however the timing was delayed in April (after the first semester).

[28] The grade 7 mathematics curriculum included eight units, and the Ministry developed two types of mathematics tests: one for checking student understanding by unit, and the other for checking student understanding by semester. In year 1 of this study, for the control group, the Ministry distributed both tests to schools; however, the number of copies distributed to each school was limited: only one copy of each test by unit in the mathematics curriculum in total, and 10 copies of mathematics tests by semester in total. On the other hand, in the treatment group, the Ministry distributed the necessary number of copies of tests for all students in grade 7 in the selected schools.

textbook distribution. In 2019 (year 2 of this study), the Ministry provided a set of teaching and learning materials, including workbooks for both groups, and organized a mutual review meeting of teachers.[29,30]

3.2.2 Assessment of math learning

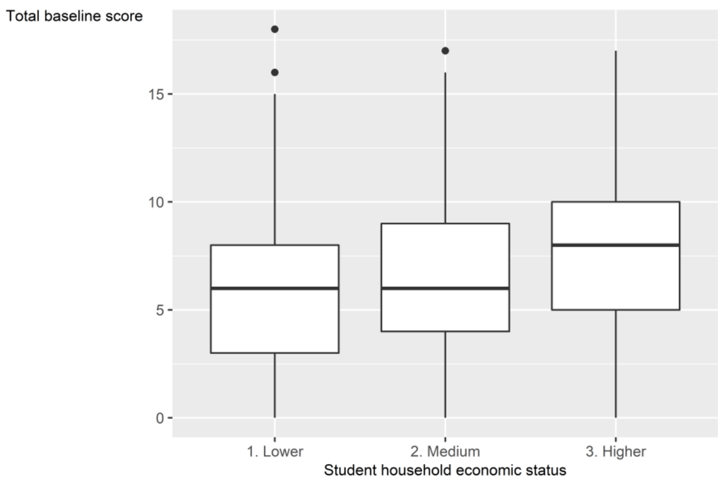
To assess their math learning, students completed written tests in all three rounds of surveys.[31] Survey teams administered the tests without the presence of teachers in the class room. To account for the progress following the curriculum, the test items differ across the three tests. The test items at the baseline survey were from math content in the primary education. The test items at the end-line were from math content in grade 7. The number of test items at the baseline and end-line surveys was respectively twenty in total, which included the problems posed in texts. The test items at the follow-up survey were mainly from math content in grade 8. The test at the follow-up survey was composed of 25 different items, of which five were the same from the end-line survey. The written test at each round of the surveys included the items to measure math learning by the cognitive skills (knowing, applying, and reasoning) and by the cognitive domains (number and operation, function, and geometry). The composition of test items is presented in Tables A.1 to A.3 in Appendix 6. Based on the common five math items in the end-line and the follow-up surveys, the test scores in the two rounds are linked using the item response theory. The process of linking test scores is reported in Appendix 7 of this chapter.

The baseline test results demonstrate that the grade 7 students did not master the foundations of math including the basic four operations. For example, only 42 percent of the students were able to correctly answer the item "43 – 17". Responding to the item "2×2×2", 34 percent of students confused multiplication with addition, answering "6". On the division item "612÷102," half of the students did

[29] The ESMATE project was completed at the end of June 2019. The Ministry integrated the project's activities into the plan and allocated a necessary budget to continue them.
[30] Teachers used teacher guides for grade 8, which were stored at schools, or they referred to the guide on the Ministry's website.
[31] The baseline survey was organized in January-March 2018, the end-line survey in September-October 2018, and the follow-up survey in September-October 2019.

not write any response. Through student interviews, the baseline survey collected information on different types of student household assets. A composite index of the different types of household assets was computed using the principal component analysis. The first principal component accounts for around 25 percent of the overall variance of the variables on different types of household assets, and all their loading coefficients are positive. Therefore, the first principal component was taken as the index of student household economic status. According to the first principal component, whole samples of students were divided into one of the three levels: lower; medium; and higher household economic status. The details of the principal component analysis are presented in Appendix 8. As shown in Figure 6, student baseline test scores modestly correlate with household economic status.

Figure 6. Boxplots of the baseline test scores by the student household economic status



Note: Data source is the baseline survey of this research. N. of students in the lower household economic status is 534 (treatment group) and 488 (control group). N. of students in the medium household economic status is 1,317 (treatment group) and 1,279 (control group). N. of students in the higher household economic status is 472 (treatment group) and 437 (control group).

3.2.3 Sampling

This research originally targeted both the primary and lower secondary education levels, focusing on 2nd grade in primary and 7th grade in lower secondary education. Considering the difference in the content of interventions and the educational level, this chapter focuses on lower secondary. As explained in Chapter 2, among 14 departments, Cabañas, La Union, San Miguel and San

Vicente, situated in the central and eastern parts of the country, were selected, based on the educational statistics. Within the four departments, there were 612 basic education public schools offering lower secondary education.[32]

Since this research targets both primary and lower secondary education levels, 6 schools that did not have the primary education level were excluded. The country suffers from security problems due to the presence of gang members inherited from past civil conflicts. Intentional homicides per 100,000 were 61.8 in 2017 (World Bank, 2021), the highest rate in the world. The schools were also affected by activities of gang members (USAID, 2017). The schools located in an area severely affected by such activities, and those that were physically difficult to access were excluded from the sampling frame. Outside our experimental design, the Millennium Challenge Corporation (MCC) planned to distribute the ESMATE textbooks in 2018. The schools targeted by MCC were also excluded from our evaluation framework. As a result, the sampling frame was comprised of 369 basic education public schools. Of these, 250 basic education public schools were randomly sampled, with half randomly assigned to the treatment group and the other half to the control group (Table 7). Stratification variables in the randomization are department and urban status. If there were several classes of the targeted grades in the sampled school, one class was randomly selected. In the baseline survey, seven schools in the treatment group and four schools in the control group were excluded due to security reasons (Koei Research & Consulting Inc. 2018).[33] In addition to these eleven schools, there were no enrolled students in grade 7 at two schools in the control group.

Based on the educational census data collected by the Ministry of Education in the country, the characteristics of 612 basic education public schools in the four departments is compared to the sampling frame (Table D.1 in Appendix 9). Because data from some schools are not included in the educational census survey data, the number of schools in Column (A) to (C) in Table D.1 does not

[32] As explained in Chapter 2, public schools in the country are called "basic education public schools" and can include preschool, primary, and lower and upper secondary levels according to the local educational needs.
[33] At the end-line survey, three schools in the control group were additionally excluded because of security reasons (Koei Research & Consulting Inc., 2019). In the other two schools, the number of 7th grade students enrolled at the baseline survey was small, and all the students were absent at the end-line survey.

exactly match with that in Table 7. The sampling frame represents well the original 612 basic education public schools in the four departments. The characteristics of the original sample of 250 schools and the remaining 237 schools after attrition of 13 schools were also equivalent between the two groups.

Table 7. Sampling frame of schools in the four departments

	Cabañas	La Union	San Miguel	San Vicente	Total
(A) N. of public schools (lower secondary)	105	146	248	113	612
(B) Schools with both cycle 1 and cycle 3 in (A)	104	144	247	111	606
(C-1) Schools without difficulty in access or security in (B)	64	68	164	105	401
(C-2) Schools not targeted by the MCC program (Sampling frame)	64	49	151	105	369
(D-1) Sampled schools (Treatment)	22	16	51	36	125
(D-2) Sampled schools (Control)	21	17	51	36	125

3.2.4 Comparison of student, teacher and school characteristics

The student, teacher, and school characteristics of the 237 schools are presented in Tables 8-1 and 8-2. The percentage of teachers who finished teacher training courses in the control group was higher than the treatment group by 10 percentage points (Table 8-1). The percentage of teachers teaching subjects other than mathematics was larger in in the treatment group than the control group by 9 percentage points. In terms of the school characteristics, the number of 7th grade students in the treatment group is larger than the control group by 9 students.[34] The differences are statistically significant at the 10 percent level. While those several differences are noted in teacher and school characteristics, the overall characteristics of treatment and the control groups are well balanced.

Because of logistical reasons, it was not possible to conduct the baseline survey before any component of the intervention package started. The baseline survey was started in mid-January 2018, just after the intervention in the treatment group began, and finished on March 1, 2018. The surveys of

[34] The difference in the number of 7th grade students is a result of the number of classes of 7th grade in the sampled schools. Class size of the sampled class is equivalent between the treatment group (average: 24 students) and the control group (average 23 students).

the treatment and control schools were conducted in parallel. As presented in Table 8-2, the difference in the baseline scores is nearly zero and not statistically significant (p-value: 0.94), which indicates that the baseline scores are well balanced. Density curves of the baseline scores are presented in Figure 7. All test items at the baseline survey were from the contents of primary education. Since the 7th grade ESMATE textbook does not include a review section of content learned in primary education and starts from new math content, it is plausible to think that the distribution of ESMATE textbooks at the beginning of the school year did not have an impact on the baseline scores in the treatment group.

Table 8-1. Comparison of characteristics of teachers and schools (baseline survey)

Content	Treatment (1)	Control (2)	Mean Diff. (3): (1)-(2)
Teacher characteristics			
Sex (Male) (%)	52.5	55.5	-2.9
Age	39.3	40.0	-0.6
Standard deviation	(10.1)	(10.1)	
Total teaching period (years)	15.6	16.3	-0.6
Standard deviation	(9.3)	(9.6)	
Academic Degree			
High school (%)	2.5	1.7	0.9
Professorate (%)	60.2	64.7	-4.5
Bachelor (%)	34.8	31.9	2.8
Master (%)	2.5	0.8	1.7
Doctor (%)	0.0	0.0	0.0
Teacher qualification (1)			
Pedagogical Bachelor (%)	3.4	3.4	0.00
Professor (%)	75.4	84.9	-9.5 *
License in Education (%)	18.7	14.3	4.4
Master's in Education (%)	1.7	0.8	0.9
Doctorate in Education (%)	0.0	0.0	0.0
Pedagogical Training Course (%)	8.5	4.2	4.3
Teacher qualification (2)			
Basic Education Teacher (Cycle I and II) (%)	15.3	15.1	0.1
Math Specialty Teacher (Cycle III and High School) (%)	60.2	58.8	1.3
Specialized in other than math (Cycle III and High School) (%)	17.8	22.7	-4.9
School characteristics			
Number of students			
N. of Student (7th grade) Morning Shift	39.7	30.6	9.2 *
Standard deviation	(25.9)	(20.7)	
N. of Student (7th grade) Afternoon Shift	20.9	20.6	0.2
Standard deviation	(11.4)	(10.2)	
N. of Student (Total)	257.6	238.8	18.8
Standard deviation	(231.9)	(176.3)	
Number of teachers	10.6	10.0	0.6
Standard deviation	(9.2)	(6.8)	
Repetition and dropout rate			
Repetition rate (morning shift of 7th grade in 2017) (%)	4.3	3.8	0.5
Repetition rate (afternoon shift of 7th grade in 2017) (%)	4.4	5.1	-0.6
Dropout rate (morning shift of 7th grade in 2017) (%)	8.9	9.4	-0.5
Dropout rate (afternoon shift of 7th grade in 2017) (%)	10.8	9.8	1.0
School facility			
Electricity (%)	100	100	0.0
Drinking Water (%)	79.7	86.6	-6.9
Computer (%)	94.9	93.3	1.6
Library (%)	25.4	25.2	0.2
Laboratory (%)	11.9	6.7	5.1
Donor support within 5 years (except ESMATE) (%)	91.5	94.1	-2.6
N. of schools	118	119	

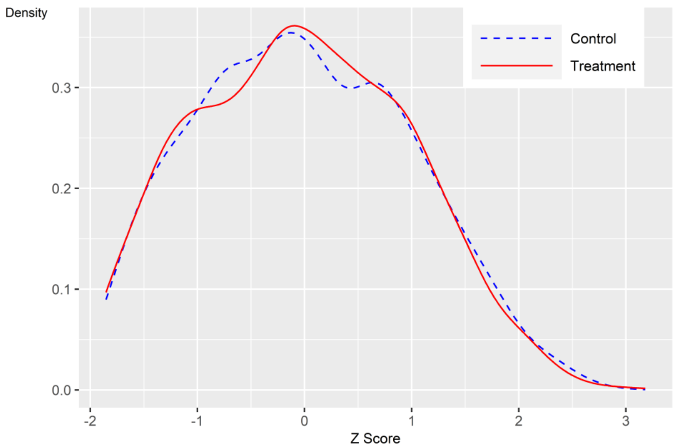
Note:
(1) Data source is the baseline survey of this research. 10% significance: *. (2) Column 3 reports the difference between the treatment and control groups, and the result of the test for the difference between the two groups. For binary values, the chi-squares test with stratified data (department and urban status) is conducted. For total teaching period (years), age, the number of students, repetition rate, and dropout rate, the Wilcoxon rank sum test with stratified data (department and urban status) is employed.

Table 8-2. Comparison of characteristics of students (baseline survey)

Content	Treatment (1)	Control (2)	Mean Diff. (3): (1)-(2)
Morning Shift (%)	58.33	63.65	-5.32
Age	13.03	12.98	0.05
Standard deviation	(1.30)	(1.28)	
Sex (Male) (%)	48.82	50.09	-1.30
Repeated (%)	22.85	21.92	0.94
N. elder brother/sister	1.76	1.79	-0.03
Standard deviation	(2.09)	(2.02)	
N. younger brother/sister	1.13	1.09	0.02
Standard deviation	(1.15)	(1.21)	
Raw test scores (total points: 20)	6.58	6.64	-0.06
Standard deviations of raw test scores	(3.55)	(3.58)	
Standardized test scores	-0.012	0.004	-0.016
Asset of study			
Math textbook 2017 (%)	25.74	26.67	-0.92
Math notebook 2017 (%)	90.40	87.63	2.78
Notebook only for Math 2017 (%)	89.24	85.47	3.76
Study desk at home (%)	95.05	95.39	0.34
Asset of student household			
Smartphone (%)	86.48	85.39	1.09
Computer (%)	29.79	29.45	0.34
Refrigerator (%)	85.36	84.25	1.12
Car (%)	30.22	29.86	0.36
TV (%)	93.37	93.79	-0.42
Tap water (%)	77.18	80.09	-2.91
Electricity (%)	96.43	96.21	0.22
Flush Toilet (%)	52.35	55.71	-3.36
Using wood for cooking (%)	61.08	62.32	-1.24
Student household economic status index	-0.005	0.006	-0.011
N. of schools	118	119	
N. of students	2, 323	2,190	

Note:
(1) Data source is the baseline survey of this research.
(2) Baseline test scores are standardized by the mean and standard deviations of baseline scores of the control group. Student household economic status index is the first component in the principal component analysis on different types of household assets. For more detail, please refer to Appendix 8 of this chapter.
(3) Column 3 reports the difference between the treatment and control groups and the result of the test for the difference in means between the two groups. The test controls for strata fixed effects constructed by the stratification variables in the random assignment (department and urban status). Robust standard errors are clustered at the school level. None of the values in this table are statistically significant.

Figure 7: Density curves of the baseline test scores (standardized scores)



Note: Data source is the baseline survey of this research.

3.2.5 Student attrition

The same students were tracked through three rounds of surveys for two years. Among 4,515 students who were present at the baseline survey, 894 were absent at the end-line survey, and 1,370 were absent at the follow-up survey. The attrition rate was 19.4 percent at the end-line survey and 29.1 percent at the follow-up survey in the treatment group. In the control group, the attrition rate was 20.3 percent at the end-line survey, and 31.6 percent at the follow-up survey. The major reasons behind the attrition were students moving to the other schools or dropping out. At the end-line survey, 11 percent of students in the treatment group and 9 percent in the control group moved to another school or dropped out. In the following year, 12 percent in the treatment group and 11 percent in the control group moved to another school or dropped out.

To check whether differential attrition occurred between the two groups, the student attrition dummy is regressed on the treatment assignment, student characteristics (baseline), and strata fixed effects constructed by department and urban status. The results are shown in Table E.1 in Appendix 10. The attrition was not differential between the two groups. The attrition occurred more among boys, students with lower baseline scores, and older students. There was also higher student attrition among those who have more younger siblings.

Grade 7 students who repeated the same grade in year 2 of this research remained in the sample and were assigned the same math test taken by the grade 8. In the follow-up survey data, 1.9 percent of students in the treatment group and 2.3 percent of students in the control group repeated 7th grade.

3.3. Impacts on student math learning

3.3.1 Estimation strategy

The impacts of the additional components in the ESMATE program on student math learning in year 1 and year 2 are estimated using following equation.

$$Y_{ijr}=\alpha_r+\gamma_r(Y_{ij0}\times Year_r)+\delta_r(Treatment_j\times Year_r)+ C_{ijr}\beta_c+ P_{kjr}\beta_p+ S_{jr}\beta_s+ D_j\beta_D+\varepsilon_{ijr} \quad (2)$$

In equation (2), Y_{ijr} in the left-hand side (r=1: end-line and r=2: follow-up) represents the math test score for student *i* in school *j* at the end-line or follow-up, and Y_{ij0} in the right-hand side is her/his test scores at the baseline survey. Baseline test scores are standardized by the mean and standard deviations of test scores of students in the control group. Based on the common five math items in the end-line and follow-up surveys, the test scores in the left-side are linked using the item response theory (IRT). The linked scores are called "IRT scores." The IRT scores are standardized by the mean and standard deviations of the scores in the control group at the end-line survey. To evaluate the impact by the cognitive skills (i.e., knowing, applying and reasoning) and domains (i.e., number & operation, function and geometry), sub-totals of test scores are also used instead of the IRT scores.

$Year_r$ is a vector of dummy variables of year *r* (r=1: end-line and r=2: follow-up). C_{ijr} is a vector of characteristics of student *i* at school *j* such as age, gender, shift at school (morning or afternoon), and the number of brothers and sisters, and characteristics of the family of student *i* such as household asset at the baseline. C_{ikr} includes the shift at school (morning or afternoon) during year 1 and year 2, and whether student *i* repeated the 7th grade in year 2. P_{kjr} is a vector of characteristics of teacher *k* at school *j*, who teaches mathematics to student *i* in year *r*, such as age, gender, and educational qualification. S_{jr}

is a vector of characteristics of school *j* such as school infrastructure and the total number of students attending school *j* at the baseline. S_{jr} includes characteristics of school principal at year 1 and 2. D_j is a vector of strata fixed effects constructed by department and urban status of school *j*. Robust standard errors are clustered at the school level. In equation (2), ε_{ijr} is an error term.

To analyze the heterogeneous impacts by the baseline scores and the student household economic status, equation (2) is expanded by adding the interaction term of the baseline score or student household economic status index and the treatment assignment as

$$Y_{ijr}=\alpha_r+\gamma_r(Y_{ij0}\times Year_r)+(\delta_{rA}+\delta_{rB}X_{ij0})(Treatment_j\times Year_r)+C_{ijr}\beta_c+P_{kjr}\beta_p+S_{jr}\beta_s+D_j\beta_D+\varepsilon_{ijr} \quad (3).$$

In equation (3), X_{ij0} stands for either a vector of the baseline scores or student household economic status index.

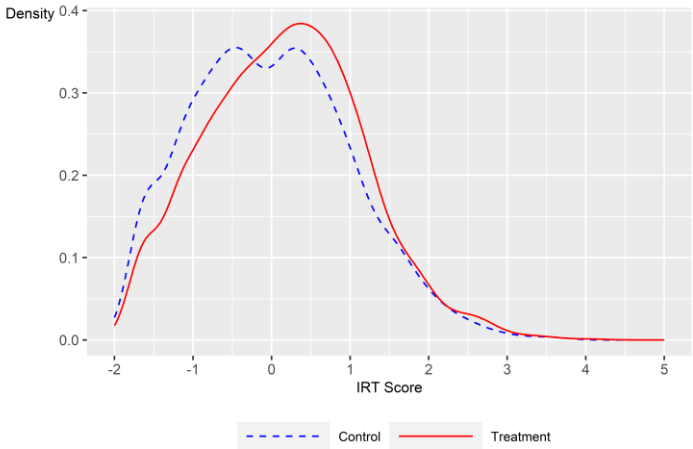
3.3.2 The impacts of the additional components in the ESMATE program

Density curves of the test scores at the end-line and follow-up surveys are presented in Figures 8 and 9. Since the test scores are linked by the item response theory, the test scores are comparable across the end-line and follow-up surveys. While the IRT scores in the treatment group were higher than the control group at the end of year 1 (Figure 8), both groups improved math learning in the following year; then, the difference between the two groups has almost disappeared at the end of year 2 (Figure 9). The average impact of the additional components in the ESMATE program on student math learning is estimated by equation (2), pooling the samples of students who took math tests both at the baseline and end-line surveys, and those who took math tests both at the baseline and follow-up surveys. The regression results are shown in Column (I)-1 through (I)-3 in Table 9. The average one-year impact of additional components in the ESMATE program is estimated at 0.17 standard deviations of the IRT scores, which is statistically significant at the 1 percent level.[35] On the other hand, the accumulated

[35] I conducted the cost-effective analysis, following the methodology presented by J-PAL (Bhula, R. et al. 2020; Dhaliwal et al. 2014). The cost-effectiveness is measured as the ratio of the aggregated impact of the project (the average impact on student learning per student multiplied by the number of students impacted) to the aggregated

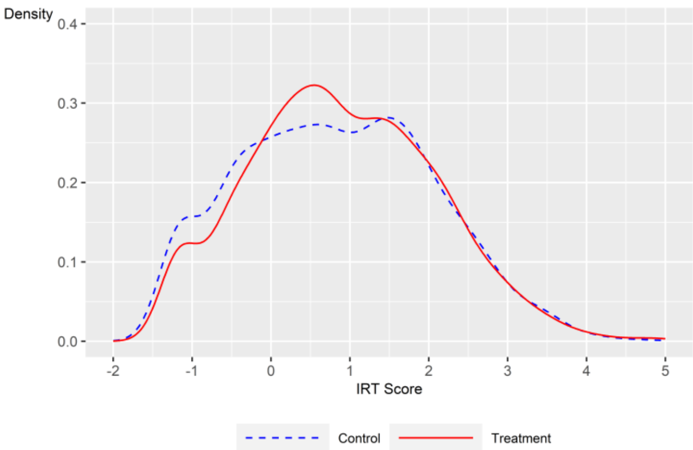
impact of the first-year intervention in the following year is estimated at 0.07, which is not statistically significant.[36] The average impacts are also estimated, using the samples of students who took math tests at all three rounds of survey. The estimated impacts become slightly higher than the estimates with the pooled sample but remain almost at the same levels (Column (II)-3 in Table 9).

Figure 8. Density curves of the end-line test scores (IRT scores)



Note: Data source is the end-line surveys of this research.

Figure 9. Density curves of the follow-up test scores (IRT scores)



Note: Data source is the follow-up surveys of this research.

cost of implementing the project. The cost-effectiveness is presented as the total standard deviations gained across the sample per 100 USD spent. The cost-effectiveness of the additional interventions of the ESMATE program, the total standard deviations gained across the sample per 100 USD spent, is estimated to be 2.68. The level of cost-effectiveness of the additional interventions in the ESMATE program is comparable to the other programs cited in Kremer et al. (2013).

[36] I tested a hypothesis that the coefficient of treatment assignment and first-year dummy was equal to the coefficient of treatment assignment and second-year dummy. The hypothesis was not rejected at the 10 percent level (p-value was 0.256).

Table 9. The average impact of the ESMATE program on student math learning

	(I)-1: Full sample	(I)-2: Full sample	(I)-3: Full sample	(II)-1: Balanced panel sample	(II)-2: Balanced panel sample	(II)-3: Balanced panel sample
Treatment×1st year dummy	0.187**	0.188***	0.172***	0.183**	0.195***	0.189***
	(0.075)	(0.058)	(0.055)	(0.080)	(0.061)	(0.060)
Treatment×2nd year dummy	0.069	0.088	0.076	0.077	0.090	0.082
	(0.096)	(0.078)	(0.072)	(0.098)	(0.080)	(0.075)
2nd year dummy	0.845***	0.844***	0.564***	0.829***	0.853***	0.609***
	(0.058)	(0.057)	(0.139)	(0.060)	(0.060)	(0.149)
Z baseline scores		0.517***	0.469***		0.523***	0.478***
×1st year dummy		(0.022)	(0.023)		(0.024)	(0.026)
Z baseline scores		0.550***	0.513***		0.552***	0.515***
×2nd year dummy		(0.025)	(0.024)		(0.025)	(0.024)
Student characteristics	No	No	Yes	No	No	Yes
Teacher characteristics	No	No	Yes	No	No	Yes
School characteristics	No	No	Yes	No	No	Yes
N. obs.	6,763	6,763	6,763	5,718	5,718	5,718
N. Clusters	232	232	232	229	229	229

***p < 0.01; **p < 0.05; *p < 0.1

(1) The data source is the three rounds of survey of this research.

(2) Robust standard errors are clustered at the school level and are in parenthesis. Strata fixed effects are controlled in all the regressions but are not shown.

(3) The IRT scores (end-line and follow-up) are standardized by the mean and standard deviations of the IRT scores of the control group in the end-line survey.

(4) Baseline test scores are standardized by the mean and standard deviations of the test scores of the control group.

To deepen the analysis of the accumulated impact, I conducted a quantile regression by regressing the IRT scores in year 1 and year 2 of the research respectively on the treatment assignment dummy. The plots of quantile treatment effect estimates are presented in Figures F-4 and F-5 in Appendix 11. The estimates of the one-year impact are positive and statistically significant for students in the treatment group from the lowest to approximately 90 percent in the distribution of IRT scores (Figure F-4). On the other hand, the estimates of the accumulated impact are positive and statistically significant for students in the treatment group from the lowest to approximately 30 percent in the distribution of IRT test scores (Figure F-5). The results also indicate that the students who improved math learning in year 1 could not advance their learning well in the following year.

To analyze the impacts of the additional interventions in the ESMATE program from a different angle, sub-totals of test scores by the cognitive domains and skills are used instead of the IRT scores. The subtotals of test scores are standardized by the mean and standard deviations of the control group at each round of surveys. The average impacts by the cognitive skills and domains are estimated using the following equation,

$$Y_{ijr}=\alpha_r+\gamma_r Y_{ij0}+\delta_r Treatment_j+C_{ijr}\beta_c+ P_{kjr}\beta_p+ S_{jr}\beta_s+ D_j\beta_D+\varepsilon_{ijr} \quad (4),$$

where subscript r represents the end-line (r=1) or the follow-up (r=2). Without IRT, we cannot directly compare test scores in the end-line and the follow-up. Thus, unlike equations (1) to (3), which are to be estimated by pooling both the end-line (r=1) and the follow-up (r=2) samples, equation (4) is estimated separately for the end-line and the follow-up.

The regression results are presented in Table 10. In year 1, while the impacts in the domains of “number and operation” and “function” are positive and statistically significant, the impact in the domain of geometry is not statistically significant. In the former two domains, students in the treatment group correctly responded to the items such as simple operations with positive and negative numbers, and direct and inverse proportions than the control group. Looking at the cognitive skills, while the impact on the knowing skill is statistically significant, the impact on the applying and reasoning skills is not. On the other hand, in year 2, all the estimated values are nearly zero and not statistically significant (Table 10).

As noted in the section 3.2.1, the distribution of ESMATE textbooks and teachers' guidebooks was delayed by around one month in the control group in year 1. While the curriculum covered in the 2018 school year was limited in the control group due to the delay, approximately the same percentage of schools completed the units No. 1 through 4 in the curriculum. [37]

The impact of the additional components in the ESMATE program is estimated, taking the sub-total of the end-line test scores corresponding to the unit 1 through 4, using equation (4) for the robustness check. The estimated one-year impact of the additional components is around 0.18 standard deviations, which remain at almost the same level as the estimates in the IRT scores in Table 10.

Table 10. The impacts of the additional components in the ESMATE program by the cognitive domains and the cognitive skills

	(I)-1 Number & Operation	(I)-2 Function	(I)-3 Geometry	(II)-1 Knowing	(II)-1 Applying & reasoning
(A) One-year impact					
Treatment	0.169***	0.182***	0.079	0.187***	0.065
	(0.057)	(0.054)	(0.063)	(0.055)	(0.059)
Num. obs.	3,619	3,619	3,619	3,619	3,619
N. Clusters	232	232	232	232	232
(B) Accumulated impact of the first-year intervention in the following year					
Treatment	-0.014	0.034	0.027	0.029	-0.057
	(0.062)	(0.049)	(0.066)	(0.058)	(0.047)
Num. obs.	3,143	3,143	3,143	3,143	3,143
N. Clusters	229	229	229	229	229

***$p < 0.01$
(1) Data source are the baseline and end-line surveys of this research.
(2) Robust standard errors are clustered at the school level and are in parenthesis.
(3) Student test scores are standardized by the mean and standard deviations of test scores in the control group at each round of survey. The standardized baseline test scores are controlled in all the regressions but not shown.
(4) Student, teacher, school characteristics, and strata fixed effects constructed by department and urban status are controlled in all the regressions but not shown.

[37] In both groups, around 90 percent of schools finished the unit No. 1 through 4 among eight units in the curriculum at the end-line survey.

3.3.3 Heterogeneity in the one-year impact of the additional components

The heterogeneity of the average treatment effect by the baseline scores and the household economic status is investigated by equations (3). The results are presented in Table 11. The heterogeneous one-year impact by student economic household status is estimated at -0.04, which is statistically significant at the five percent level (Column (I)). On the other hand, the impact of the additional treatment is not heterogeneous by the baseline test scores (Column (II)). To analyze the heterogeneous impact by student household economic status from a different angle, student household economic level dummies (low, medium, or higher status) are used instead of the status index in equation (3).[38] The regression results are shown in Table F.1 in Appendix 11 of this chapter. The net impact for higher economic status is close to zero (Column (I) and (III) in Table F.1).[39] Since the impacts for students with the lower and median economic status are positive and statistically significant, the additional components reduced the disparity in learning outcomes by the student household economic status.

As seen in section 3.2.2, the baseline test scores and the household economic status level are correlated with each other; thus, both of the interaction terms with the treatment assignment are included in equation (3). As shown in Column (III) in Table 11, the estimated values of heterogeneous impact by student household economic status remain at almost the same level, even after controlling for the heterogeneous impact by the baseline scores. Furthermore, the triple interaction term of the treatment assignment, student household economic status index, and baseline test scores, and the interaction term of student household economic status index and baseline test scores are added to the equation. The coefficient of the triple interaction term is -0.05, which is statistically significant at the one percent level (Column (IV) in Table 11). Students in the control group with higher baseline scores and household

[38] Please refer to Appendix 8 on the details of the category of low, medium and high household economic status.
[39] Additionally, I investigated the heterogeneous impact on the frequency of homework assignments by student household economic status using student household economic level dummies (low, medium, or higher status). The results are presented in Table F.2 in Appendix 11. The impact is larger for students with lower economic status.

economic status obtained higher test scores at the end-line survey. The additional components in the ESMATE program reduced the disparity in learning outcomes.

Table 11. Heterogeneity in the impact of the additional components in the ESMATE program

	(1)	(2)	(3)	(4)
Treatment × 1st year dummy	0.172***	0.168***	0.170***	0.184***
	(0.055)	(0.055)	(0.055)	(0.055)
Treatment × 2nd year dummy	0.073	0.071	0.071	0.069
	(0.072)	(0.073)	(0.072)	(0.072)
2nd year dummy	0.595***	0.594***	0.594***	0.599***
	(0.140)	(0.140)	(0.140)	(0.139)
Treatment × 1st year dummy	-0.048**		-0.046*	-0.053**
× Student economic status index	(0.024)		(0.024)	(0.024)
Treatment × 2nd year dummy	-0.017		-0.014	-0.015
× Student economic status index	(0.033)		(0.032)	(0.033)
Treatment × 1st year dummy		-0.039	-0.030	-0.023
× Z score baseline		(0.043)	(0.042)	(0.041)
Treatment × 2nd year dummy		-0.039	-0.037	-0.037
× Z score baseline		(0.048)	(0.047)	(0.047)
Treatment × 1st year dummy				-0.060***
× Z score baseline				(0.021)
× Student economic status index				
Treatment × 2nd year dummy				-0.005
× Z score baseline				(0.026)
× Student economic status index				
Student economic status index	0.066***	0.042***	0.065***	0.069***
× 1st year dummy	(0.014)	(0.012)	(0.014)	(0.014)
Student economic status index	0.042*	0.034**	0.041*	0.042**
× 2nd year dummy	(0.023)	(0.015)	(0.022)	(0.023)
Z score baseline	0.471***	0.491***	0.487***	0.484***
× 1st year dummy	(0.022)	(0.029)	(0.028)	(0.028)
Z score baseline	0.518***	0.538***	0.537***	0.536***
× 2nd year dummy	(0.024)	(0.035)	(0.035)	(0.035)
Z score baseline × 1st year dummy				0.030***
× Student economic status index				(0.010)
Z score baseline × 2nd year dummy				0.014
× Student economic status index				(0.019)
Num. obs.	6,763	6,763	6,763	6,763
N. Clusters	232	232	232	232

***p < 0.01; **p < 0.05; *p < 0.1
(1) Data source are the baseline and end-line surveys of this research.
(2) Robust standard errors are clustered at the school level and are in parenthesis.
(3) The IRT scores (end-line and follow-up) are standardized by the mean and standard deviations of the IRT scores of the control group at the end-line survey. Baseline test scores are standardized by the mean and standard deviations of the test scores of the control group.
(4) Student, school, teacher (end-line) characteristics, and strata fixed effects constructed by department and urban status are controlled but not shown. The student economic status index is the first principal component. Refer to Appendix 8 for the details.

3.4. Causal mediation analysis of the first-year intervention

3.4.1 Mediators on the impact of the first-year intervention

As presented in section 3.3.1, the additional components improved student math learning by 0.17 standard deviations of the IRT scores. This section analyzes the mechanism of the additional components in year 1, beginning with a review of the process data. The additional intervention in the ESMATE program tried to increase the length of time when students engaged in math learning in a lesson. The survey of this research conducted math lesson observations at the end-line survey, which measured the length of time when students engaged in math learning. Surveyors counted the length of time throughout a lesson when half of students in a class either solved a math problem, referred to their notebook or textbook, or consulted with each other. The percentage of the engaged time in a lesson was slightly longer in the treatment group (22.4 percent) than the control group (18.9 percent).[40]

The additional intervention also aimed at improving the management of the annual teaching plan, and facilitate teachers to provide additional math classes when the plan was delayed. At the beginning of school year, teachers in the treatment group developed an annual math teaching plan, which they regularly reviewed during periodical teacher meetings. The annual math teaching plan is a simple one-page format with a year-long calendar that defines which page of textbook is to be taught on which day. During the inter-semestrial review meeting of teachers, they checked their progress and revised it considering the remaining number of lessons in the school year. The process might have facilitated teachers to identify the delay of curriculum and provide additional math class. The percentage of teachers who provided additional math class was larger in the treatment group (52.5 percent) than the control group (37.7 percent).

Furthermore, the distribution of math workbooks was intended to support student study math at home. Since teachers can use math workbooks to assign students homework regularly, the distribution

[40] The lesson observation also checked the following instructional routine of teachers: (a) teacher posed math problem; (b) teacher checked student notebooks; (c) teacher walked around the classroom to check notebooks; (d) teacher advised students to consult with each other; (e) teacher told students to check their answers; and (f) teacher instructed students to try incorrectly solved problems again. The surveyors noted these points according to the teachers' speech and/or behaviors.

of math workbooks would increase the frequency of homework assignment. The percentage of teachers who assigned math homework four or more times in a week was larger in the treatment group (69.5 percent) than the control group (52.6 percent). On the other hand, the self-checks of answers in the workbook by students were not frequent. At the follow-up survey, the survey team randomly selected up to 3 students in grade 8 and checked their workbooks in the previous grade. On average, in around half of the pages that students worked on, answers were not self-checked by students. The status of homework was not also well checked by teachers. More than half of the pages that students worked on were not checked by teachers.

The analysis on the process data suggests several possible mediators for the improvement of student math learning in year 1, i.e., the volume of engaged time in a math lesson, the provision of additional math class, and the frequency of homework assignments.[41] The next sections explain the estimation strategy of the causal mediation effect and discuss the results.

3.4.2 Estimation strategy of the causal mediation effect

Causal mediation analysis employing the method of Imai et al. (2010) and taking the abovementioned three possible mediators sheds light on the mechanism underlying the impact on student math learning in year 1. Since the treatment group received a set of additional components, a simple strategy of using the treatment assignment as an instrument to estimate the impact via respective mediator violates the exclusion restriction assumption.

Following Imai et al. (2010), the conceptual framework of the causal mediation analysis is introduced. The mediation effect, λ_i, for student *i* can be described as

$$\lambda_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0)),$$

[41] The additional components in the ESMATE program for the treatment group include initial on-site advice for school principals, which intended to increase the frequency of their lesson observations and suggestions for teachers. The survey conducted interviews with teachers about math teaching and support received by the school principal. There was no difference between the two groups in the support by school principal, i.e., frequencies of lesson observations and suggestions provided during those observations.

where t represents the treatment assignment dummy that takes 0 or 1, and $M_i(t)$ is a potential value of mediator. The causal mediation effect $\lambda_i(t)$ is a change in Y_i due to the change in $M_i(0)$ from $M_i(1)$, holding the treatment status of student i constant at t. Then, the direct effect, ζ_i, for student i is defined as

$$\zeta_i(t) \equiv Y_i(1, M_i(t)) - Y_i(0, M_i(t)).$$

The direct effect represents the treatment impact that is not brought via the mediator $M_i(t)$. The average treatment impact can be decomposed to the average causal mediation effect and average causal direct effect.

Imai et al. (2010) introduces the following assumptions, called “Sequential Ignorability (SI),” to identify the average causal mediation effect;

$$\text{SI assumption (1): } Y_i\{(t, m), M_i(t)\} \perp T_i \mid X_i = x, \text{ and}$$

$$\text{SI assumption (2): } Y_i(t, m) \perp M_i(t) \mid T_i = t, X_i = x.$$

SI assumption (1) indicates that the treatment assignment is independent of the potential mediator and potential outcome. The assumption is satisfied in a randomized controlled trial, since the treatment assignment T_i is random. On the other hand, SI assumption (2) requires that the mediator is independent of the potential outcome, given the individual treatment assignment and pre-treatment covariates (X_i). The assumption is strong since it will be violated once there are any unobservable pre-treatment covariates or post-treatment covariates that affect both the mediator and outcome variable.

Under SI assumptions, the average causal mediation effect is estimated using the following two equations,

$$M_{ij} = \alpha^1 + \varphi^1 Y_{ij0} + \tau^1 Treatment_j + C_{ij}\xi_c^1 + P_{kj0}\xi_p^1 + S_k\xi_s^1 + D_k\xi_D^1 + \omega_{ij}^1 \text{ (6-1), and}$$

$$Y_{ij} = \alpha^2 + \tau^2 Treatment_j + \theta^2 M_{ij} + \varphi^2 Y_{ij0} + P_{kj0}\xi_p^2 + S_k\xi_s^2 + D_k\xi_D^2 + C_{ij}\xi_c^2 + \omega_{ij}^2 \text{ (6-2).}$$

where M_{ij} is a mediator, i.e., the engaged time in a math lesson; the frequency of math homework assignments; and the provision of supplementary math classes. The average causal mediation effect is expressed as the product of coefficients $\hat{\tau}^1$ and $\hat{\theta}$.

Since the impact of a mediator on an outcome variable might vary between the treatment and the control groups, equation (6-2) is expanded by adding the interaction term of the treatment assignment and the mediator as,

$$Y_{ij}=\alpha^3+\varphi^3 Y_{ij0}+\tau^3 Treatment_j+\theta^3 M_{ij}+\eta^3 (M_{ij}*Treatment_j)+P_{kj0}\xi_p{}^3+S_k\xi_s{}^3+D_k\xi_D{}^3+C_{ij}\xi_c{}^3+\omega_{ij}{}^3 \quad (6\text{-}3).$$

Then, the heterogeneous causal mediation effect by the treatment group is expressed as the product of coefficients $\hat{\tau}^1$ and $(\hat{\theta}+\hat{\eta})$.[42]

Since the unobserved pre-treatment confounders of teachers or school principals might have affected both mediator and student math learning outcomes, the sensitivity analysis proposed by Imai et al. (2010) is conducted to see how the estimates changes when the assumption does not hold. When the SI assumptions hold, the error terms (ω) in (6-1) and (6-2) or (6-1) and (6-3) will not be correlated. The correlation of the error terms ρ is a parameter, which will be used in the sensitivity analysis. A larger value of ρ in the absolute term results in a larger bias in the estimation of the average causal mediation effect.

3.4.3 Average causal mediation effects

The results of causal mediation analysis are presented in Table 12. Among the three mediators examined, only the provision of additional math class has a positive average causal mediation effect on student math learning. The impact is estimated at 0.022 standard deviations of the IRT scores,

[42] The R package "mediation" (Tingley et al., 2014) was used to estimate the average causal mediation effects. The effects are estimated by quasi-Bayesian Monte Carlo approximation. The number of Monte Carlo draws for quasi-Bayesian approximation in this study is 10,000.

statistically significant at the 10 percent level (Column (I)). The average causal mediation effect represents around 10 percent of the total impact. When examining the heterogeneity by the treatment and the control groups, while the average causal mediation effect on the control group is nearly zero and not statistically significant, the impact on the treatment group is 0.039 standard deviations, statistically significant at the 5 percent level.

On the other hand, neither the average causal mediation effect through the engaged time in math lessons nor the frequency of homework assignments are statistically significant (Column (II) and (III)). In terms of the heterogeneity by the status of the treatment assignment, while the frequency of homework assignments had a positive average causal mediation effect for the control group (0.028 standard deviations, statistically significant at the 5 percent level), the effect is not statistically significant in the treatment group.

Table 12. Results of the causal mediation analysis

	(I) Additional math class	(II) Frequency of homework	(III) Engaged time
Panel A (OLS)			
Impact of the treatment on the mediator	0.214***	0.341***	0.040**
	(0.062)	(0.089)	(0.019)
Panel B (Mediation analysis)			
ACME (average)	0.022*	0.003	-0.011
	[-0.002; 0.050]	[-0.015; 0.020]	[-0.036; 0.010]
ADE (average)	0.178***	0.202***	0.210***
	[0.077; 0.280]	[0.137; 0.270]	[0.109; 0.310]
Total Effect		0.199***	
		(0.051)	
% of impact mediated by ACME	10.6%	1.44%	-4.48%
ACME (Treatment)	0.039**	-0.023	-0.013
	[0.004; 0.080]	[-0.051; 0.010]	[-0.041; 0.010]
ACME (Control)	0.006	0.028**	-0.009
	[-0.030; 0.040]	[0.006; 0.050]	[-0.041; 0.020]
P value on hypothesis of ACME (Treatment) = ACME(Control)	0.153	0.005	0.812
Num. obs.	3,619	3,619	3,619
N. Clusters	232	232	232

Note: (a) Data source are the baseline and end-line of this research. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$ (b) Mediators analyzed in Column (I): Provision of additional math class when curriculum was delayed; mediator analyzed in (II): Frequency of homework assignments in a week; and (III): Percentage of engaged time in a math lesson. The frequency of homework assignments was surveyed by teacher questionnaires with the following four levels (1: Never; 2: Once a week or less; 3: Two or three times a week; and 4: Four or more times a week). (c) Panel A presents the OLS estimates of the treatment impact on the mediator. Panel B shows the results of causal mediation analysis. The total effect in Panel B shows the OLS estimate of the treatment impact. Student, teacher, and school characteristics, and the strata fixed effects constructed by department and urban status are controlled in all the regressions but not shown. As the specification of regression equation is different from Table 9, the OLS estimate of total effect does not match with Column (II)-3 in Table 9. (d) Dependent variable in Panel B is the IRT scores (end-line) that are standardized by the mean and standard deviations of the scores of the control group. Baseline test scores are standardized by the mean and standard deviations of the test scores of the control group.

3.4.4 Sensitivity analysis

The estimates of the average causal mediation effect in the previous section are under the SI assumption that the correlation of error terms in equation (6-1) and (6-2), and that of (6-1) and (6-3) is respectively zero. However, unobserved characteristics of teachers might have affected both the mediator and student math learning, which could bring a bias in the estimates. Therefore, the sensitivity analysis on the estimates of the average causal mediation effect that had a statistically significant value is conducted. The results of the sensitivity analysis are presented in Figure G-1 and G-2 in Appendix 12. The sensitivity parameter is the correlation of error terms, ρ. A negative value of the correlation of error terms produces upward bias in the estimate of the average causal mediation effect, and a positive value produces downward bias. As shown in the figures, the estimates of the average causal mediation effect are sensitive to the changes in the correlation. For example, when the value of correlation is 0.10, the average causal mediation effect of the provision of additional math classes for the treatment group becomes zero.

Though the error terms in equation (6-1), (6-2), and (6-3) are not observable, the correlation of the residuals from equations (6-1) and (6-2), and those from equations (6-1) and (6-3), can be checked. Scatter plots of the residuals are presented in Figure G-3-1 to G-3-4 in Appendix 12. If the pattern of scatter plots is not random, it suggests that there would be confounders that are not captured well in the regression equations above. The correlations between the residuals are nearly zero, and the scatter plots in those figures do not show a non-random pattern.

Among the three mediators in the causal mediation analysis, one of them could be a post-treatment confounder that affects both the other mediator and student math learning. The correlation between the percentage of engaged time in math lessons and the provision of math classes in the treatment group is checked. The increase in engaged time in a math lesson was not correlated with the provision of additional math classes (the correlation value: 0.02). The provision of math classes and the frequency of math homework assignments were slightly correlated in the treatment group (the correlation value: 0.08). Logically, it is plausible to think that when additional math class was provided, homework would have been assigned in the additional math class, which resulted in the increase in the frequency of homework assignments. On the other hand, it is difficult to imagine that frequent math homework caused a delay in math curricular, which brought the provision of additional math lesson.[43]

3.5. Discussions

3.5.1 Interpretation of the results of mediation analysis

While the results of causal mediation analysis suggests that the provision of additional math class is the mediator for the improvement of math learning, the mediator can be a proxy for the shift of teacher attention during math lessons towards students with lower math achievement. At the inter-semestrial meetings, teachers in the treatment group checked the percentage of students in their classes who received low scores on the diagnostic test. The process might enable them to better understand the status of math learning among their students, and strengthen their supports in math lessons. The primary objective of the additional classes was catching up following the delay in the curriculum. When teachers focused on students with low academic achievement in daily math lessons, it might have caused a delay in teaching plans.

The average causal mediation effect accounted only for 10 percent of the total impact of the additional components. One of the possible reasons for this is the technical limit on lesson observations in the survey. At the end-line survey, surveyors counted the length of student engaged time during a lesson when half of students in the class were either solving a math problem, referring to their notebook

[43] The correlation between the provision of additional math classes and homework assignments might attenuate the impact of the mediator on student math learning outcomes. To check the robustness of the impact of the provision of additional math class on student learning outcomes, I included the other two mediators in the control variable to estimate the coefficient of the provision of math class in equation (6-2). The coefficient remained at almost same level. Without controlling the other two mediators, the coefficient of the provision of additional math classes in equation (6-2) was 0.094 (standard error: 0.047; p-value: 0.048). After including the other two mediations in equation (6-2), the coefficient of the provision of additional math classes is 0.089 (standard error: 0.047; p-value: 0.061).

or textbook, or consulting with each other. But the student engaged time survey might not have captured well the quality of teaching including the shift of teacher attention to those with lower math achievement.

3.5.2 Student math learning and household economic status

Students with lower household economic status receive less support from family for study at home in El Salvador. The follow-up survey additionally collected the information on parents' educational backgrounds through student interviews. Mothers' educational attainment positively correlates with household economic status (Figure F.1 in Appendix 11), and mothers with higher educational attainment support student study at home more frequently in a week (Figure F.2). As shown in section 3.3.3, in year 1, the additional components in the ESMATE program improved math learning for students with lower or middle economic status. The additional components in the ESMATE program reduced the disparity in learning by student household economic status. In the sampled schools in this study, there was a variety of household economic status levels among students in a school.[44] It is plausible to think that teachers do not differentiate their teaching for each student according to the household economic status level. As shown in Figure F.3, the volume of students with lower baseline scores is larger in low and medium household economic status. The additional components in the ESMATE program might have shifted the focus of teachers for students with lower baseline score group in their classes.

3.5.3 The impact of the ESMATE program on student math learning

While the additional components in the ESMATE program improved math learning in the domains of number and operation, and function in year 1, they did not work to improve learning in geometry. The result in this study is in line with the other research on a different structured pedagogy program in Costa Rica. In this case, Berlinski and Busso (2017) investigated the impact of a structured

[44] The average percentage of low economic household students per school was around 25 percent, and the standard deviation was 21 percentage point. The average percentage of medium economic household students per school was around 58 percent, and the standard deviation was 18 percentage point.

pedagogy program in mathematics at the lower secondary level. The program included a set of teacher manuals and student workbooks, and teacher training aimed at changing teaching practices to improve student math learning in geometry. While teachers and students utilized the distributed materials, the students in the control group obtained higher test scores than the treatment group.

The accumulated impact of the first-year interventions in the following year was positive; however, it was not statistically significant. The cause behind the estimated accumulated impact is observed in student understanding of linear equation that they learned in grade 7. At the end of grade 7, most students in both groups had difficulty solving simple equation problem like "solve the equation, $2x=8$" (the correct response rates: 17.8 percent in the treatment group, and 14.3 percent in the control group). Since math content in grade 8 involve equations, the low-level understanding of linear equations in the grade 7 hampered student math learning in the following year. In the follow-up survey, the correct response rates to simple simultaneous equations were less than 10 percent in both groups. The root cause of the struggle can be found in the initial learning level in grade 7. At the beginning of 7th grade, 62 percent of the students in the treatment group correctly responded to the item "answer the number in □ of equation, $2\times\square=8$". While the structure of the item looks similar to the simple equation item above, the correct response rate to the latter was significantly lower. The result indicates that students in grade 7 had difficulty bridging math learning from the primary to the lower secondary level, when math content becomes more theoretical and abstract.

3.6 Conclusions

While recent debates on educational development focus on the learning crisis in primary education, the crisis in the lower secondary education level is equally profound. Through lesson observations in several countries in Latin America, Bruns and Luque (2015) argued that poor student learning results could be directly linked to the failure of teachers to keep students engaged in learning. One of the approaches to improve student learning is the structured pedagogy program that combines different types of interventions with the provision of teaching and learning materials. A structured pedagogy program in mathematics, the ESMATE program, was developed by the Ministry of Education in El Salvador with technical cooperation from JICA. While the Ministry scaled up the ESMATE program nationwide in the lower secondary education level in 2018, the Ministry could not offer some parts of the program because of budget constraints.

This study investigated the impact of the additional components on student math learning through a randomized controlled trial. The average one-year impact of the additional components in the ESMATE program is estimated at 0.17 standard deviations of the IRT scores. While the additional components improved math learning for students with lower and medium household economic status, they did not have the impact on students with higher household economic status. The mechanism of the additional components was then studied using causal mediation analysis, taking three mediators; student engaged time in math learning during a lesson; the frequency of homework assignments, and the provision of additional math classes when the curriculum delayed. Among the three mediators, the average causal mediation effect of the provision of additional math classes is positive and statistically significant. Since the provision of additional math classes can be a proxy for the shift of attention of teachers in math lessons, the results should be interpreted with a caution.

At the second year, both groups received the ESMATE textbooks, teachers' guidebooks, and student math workbooks. The accumulated impact of the first-year intervention in the following year is estimated at nearly zero and not statistically significant. At the end of grade 7, most students in both groups had difficulty solving simple equation problems. Since math content in grade 8 often involve equations, the low-level understanding of linear equations in grade 7 hampered progress in learning in the following year. The root cause of the low achievement level is found in the beginning of the lower secondary education. Students entered lower secondary education without the foundations of mathematics, and then they struggled with math content as it becomes more theoretical and advanced. To improve math learning in lower secondary education, it is essential to increase support for students to catch up on foundational math concepts and to transition smoothly in math learning from the primary to the lower secondary level, at the early stage in the lower secondary education.

Appendix 6. Composition of math test

Appendix 7. Linking the test scores at the end-line and follow-up surveys

Appendix 8. Principal component analysis of the different types of student household assets

Appendix 9. Comparison of characteristics of schools in four departments

Appendix 10. Attrition analysis

Appendix 11. Figures and Table on the analysis of impact of the additional components in the ESMATE program

Appendix 12. Sensitivity analysis plots

Reference

Berlinski, Samuel, and Matias Busso. 2017. "Challenges in educational reform: An experiment on active learning in mathematics." *Economics Letters* 156: 172-75.

Bruns, Barbara and Javier Luque. 2015. *Great Teachers: How to Raise Student Learning in Latin America and Caribbean.* Washington DC: World Bank.

Berry, James, Harini Kannan, Shobhini Mukherji, and Marc Shotland. 2020. "Failure of frequent assessment: An evaluation of India's comprehensive evaluation program." *Journal of Development Economics*, 143, 102406.

Bhula, Radhika, Meghan Mahoney, and Kyle Murphy. 2020. Conducting cost-effectiveness analysis (CEA). https://www.povertyactionlab.org/resource/conducting-cost-effectiveness-analysis-cea.

Cooper, Harris, Jorgianne Robinson, and Erika A. Patall. 2006. "Does Homework Improve Academic Achievement? A Synthesis of Research, 1987-2003." *Review of Educational Research* 76 (1): 1-62.

de Hoyos, Rafael, Alejandro J. Ganimian, and Peter A. Holland. 2019. "Teaching with the Test: Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina." *The World Bank Economic Review* 00 (0): 1-22.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. 2014. Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education. In *Education Policy in Developing Countries*, edited by P. Glewwe, 285-338. Chicago: University of Chicago Press.

Fan, Huiyong, Jianzhong Xu, Zhihui Cai, Jinbo He, Xitao Fan. 2017. "Homework and students' achievement in math and science: A 30-year meta-analysis, 1986 – 2015." *Educational Research Review* 20: 35-54.

Hanushek, Eric A., and Ludger Woessmann. 2012. "Schooling, educational achievement, and the Latin American growth puzzle." *Journal of Development Economics* 99 (2): 497-512.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15 (4): 309-34.

Jackson, C. Kirabo, and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Economic Journal: Applied Economics*, 1:4, 85–108.

Kerwin, Jason, T., and Thornton, Rebecca L. 2020. "Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures." *Review of Economics and Statistics,* vol. 103 (2): 251-64.

Koei Research & Consulting Inc. 2018. *Impact Evaluation for the Project for the Improvement of Mathematics Teaching in Primary and Secondary Education in El Salvador: Baseline Survey Report*. JICA. Unpublished.

Koei Research & Consulting Inc. 2019. *Impact Evaluation for the Project for the Improvement of Mathematics Teaching in Primary and Secondary Education in El Salvador: End-line Survey Report*. JICA. Unpublished.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science,* 340: 297.

Marco, Gary L. 1977. "Item characteristics curve solutions to three intractable testing problems." *Journal of Educational Measurement,* 14: 139-160.

Maruyama, Takao, and Takashi Kurosaki. 2021. "Developing textbooks to improve math learning: Empirical Evidence from El Salvador." *JICA RI working paper*, No.217, JICA RI.

Ministerio de Educatión, El Salvador. 2014. *EDUCACIÓN DE EL SALVADOR EN CIFRAS 2009-2014*.

Ministerio de Educatión, El Salvador. 2016. Bases de Datos por Centros Escolares - Censo Escolar: Inicial Directorio 2016. El Salvador. https://www.mined.gob.sv/estadisticas-educativas/item/6116-bases-de-centros.html.

Ministerio de Educatión, El Salvador. 2018. Bases de Datos por Centros Escolares - Censo Escolar: Inicial Directorio 2018. El Salvador. https://www.mined.gob.sv/estadisticas-educativas/item/6116-bases-de-centros.html.

Ministerio de Educatión, El Salvador. 2018. Bases de Datos por Estudiante - Censo Escolar Inicial: Año 2018. https://www.mined.gob.sv/estadisticas-educativas/item/7154-bases-de-datos-spss.html.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *The Economic Journal*, 120 (August), 187–203.

OECD 2021. PISA 2021 Mathematics Framework. https://pisa2021-maths.oecd.org/

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt and Emmanuel Jimenez. 2016. *The Impact of education programmes on learning and school participation in low- and middle- income countries*. London: International Initiative for Impact Evaluation.

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. "mediation: R Package for Causal Mediation Analysis." *Journal of Statistical Software* 59 (5).

USAID 2017. *EVALUACIÓN DEL SECTOR EDUCACIÓN DE EL SALVADOR USAID/EL SALVADOR*.

World Bank. 2021. World development indicators. https://data.worldbank.org/indicator/

Appendix 6

Table A.1. Composition of math test (baseline)

Item No.	Item content	Cognitive domain	Cognitive skill
1	To add two-digit numbers with carrying to the position of tens	NO	Knowing
2	To subtract a two-digit number from a two-digit number with borrowing from the position of tens	NO	Knowing
3	To find the product of three 2's	NO	Knowing
4	To operate division of three-digit number by three-digit number without remainder	NO	Knowing
5	To operate multiplication and division of one-digit numbers successively	NR	Knowing
6	To subtract a number with one decimal place from a number of the same type without borrowing	NO	Knowing
7	To operate division without residue of numbers with one decimal place	NO	Knowing
8	To add two proper fractions with the same denominator without reduction	NO	Knowing
9	To find the product of two proper fractions without reduction	NO	Knowing
10	To find the quotient of two proper fractions without reduction	NO	Knowing
11	To find the least number which can be divided without remainder both by 4 and by 6	NO	Knowing
12	To find the number which gives 3 when subtracting it from 7 (the item is presented in the form of equation representing the unknown number by □)	NR	Applying
13	To find the number which gives 8 when multiplying 2 by that number (the item is presented in the form of equation representing the unknown number by □)	NR	Applying
14	To find how many times 4 is 8 (the unknown number is represented by □)	NR	Knowing
15	To find the unknown number in the table which shows the relation between the quantity of the same goods and their total price	NR	Knowing
16	To find the unknown number in the table which shows the relation between the quantity of workers and the quantity of days which is needed to finish the same work	NR	Applying
17	To draw the development view of a die (an illustration of cutting it is attached)	G	Reasoning
18	To complete the figure which shows the left-hand side of a symmetric form	G	Knowing
19	To find the area of a rectangle from the length of its base and height	QM	Knowing
20	To find the sum of the length of the sides of the rectangle of the previous item	QM	Knowing

Note: Test items are developed in this research. NO: Number and Operation; QM: Quantity and Measurement; NR: Numerical and Quantitative Relation; G: Geometry

Table A.2. Composition of math test (End-line)

Item No.	Item content	Cognitive domain	Cognitive skill
1	To write a negative number which corresponds to a point in the number line	NE	Knowing
2	To add a one-digit negative number and a one-digit positive number (the answer is positive)	NE	Knowing
3	To subtract a one-digit negative number from a one-digit positive number (the answer is a two-digit number)	NE	Knowing
4	To multiply a one-digit positive number by a one-digit negative number (the answer is a one-digit number)	NE	Knowing
5	To divide a one-digit negative number by a one-digit negative number	NE	Knowing
6	To find the greatest common divisor of three integers	NE	Knowing
7	To find the expression of the sum of two numbers, one of which is represented by the letter x	NE	Knowing
8	To add two polynomials of degree one in one variable	NE	Knowing
9	To find the value of a polynomial of degree one in one variable when the variable is substituted by a number	NE	Knowing
10	To solve an equation of the first degree (by transposition)	NE	Knowing
11	To solve an equation of the first degree (by dividing both sides by the coefficient of the variable)	NE	Knowing
12	To solve an equation of the first degree (by transposition and division)	NE	Knowing
13	To find the unit price from the sum of the money, the number of units and the change	NE	Knowing
14	To fill a blank in the table which shows the relation between the number of units and the total price knowing the unit price	F	Knowing
15	To fill a blank in the table which shows the relation between the length of base and height of rectangles of the same area	F	Knowing
16	Given the graph of proportion, to find the coefficient of the variable in the equation	F	Applying
17	To find the total number knowing the proportion and the number of one part in a pie chart	S	Applying
18	To draw a figure which is symmetric to a given triangle with respect to a line	G	Knowing
19	When two circles have two points in common, to find the angle formed by the line which connects the centers and the line which connects the common points	G	Reasoning
20	To find the number of the lateral face of a prism whose base is rectangle	G	Knowing

Note: Test items are developed in this research. NE: Number and Expression; F: Function; G: Geometry; S: Statistics

Table A.3. Composition of math test (Follow-up)

Item No.	Item content	Cognitive domain	Cognitive skill
1	To subtract a one-digit negative number from a one-digit positive number (the answer is two-digit number)	NE	Knowing
2	To add two polynomials of degree one in one variable	NE	Knowing
3	To find the value of a polynomial of degree one in one variable when the variable is substituted by a number	NE	Knowing
4	To solve an equation of the first degree (by transposition and division)	NE	Knowing
5	Given the graph of proportion, to find the coefficient of the variable in the equation	F	Applying
6	To arrange the terms of a polynomial in one variable according to their degree	NE	Knowing
7	To multiply two monomials	NE	Knowing
8	To add two polynomials of degree one in two variables	NE	Knowing
9	To solve a system of equations of the first degree in two variables	NE	Knowing
10	To solve a written problem in which two unknown numbers appear	NE	Knowing
11	To find the values of two constants so that two systems of equations of the first degree in two variables have the same solution	NE	Applying
12	To find the y-intercept of the graph of a function of the first degree	F	Knowing
13	To fill the table which represents some values of variables of a function of the first degree	F	Knowing
14	To find the rate of change of a function of the first degree from a table of the values of their variables	F	Knowing
15	Given the values of two pairs of variables and one variable of the third pair of a function of the first degree, to find the value of another variable of the third pair (the function represents the relation between the height of water and the time)	F	Knowing
16	To find the equation of the function of the first degree whose graph is symmetric to the graph of the given function with respect to the abscissa	F	Reasoning
17	To find the value of the corresponding angle to the given one formed by two parallel lines and a transversal	G	Knowing
18	To find the applicable condition for congruence of given triangles	G	Knowing
19	To find the length of one side of a parallelogram when the length of two consecutive sides are given	G	Knowing
20	To find the measure of an internal angle of a parallelogram when the measure of its opposite angle is given	G	Knowing
21	To find the measure of an angle formed by a diagonal and one side of a rectangle when the measure of an angle formed by its diagonals is given	G	Applying
22	To find the name of a solid generated by a right-angled triangle when it is rotated on one of its catheti	G	Knowing
23	To find the volume of a pyramid whose base is square from the length of the side of its base and its height	G	Knowing
24	To find the number of rods of a certain range of length from a frequency distribution table	S	Knowing
25	To find the midpoint of the modal class of a frequency distribution table	S	Knowing

Note: (1) Test items are developed in this research. NO: Number and Expression; F: Function; G: Geometry; S: Statistics. (2) No. 1 through 5 are the same items from the end-line survey.

Linking the test scores at the end-line and follow-up surveys

To estimate the values of item difficulty and item discrimination of the math items at the end-line and the follow-up surveys, the sample of students who were present at both surveys (N=2,859 students at each round of survey) was used. The two-parameter logistic model is employed to estimate the values.

The values item difficulty and item discrimination of the math items at the end-line survey are estimated, then the values at the follow-up survey are estimated. As noted in section 3.2.2 in this chapter, there are five common math items in the two rounds of surveys. The five items are used for equating the values at the follow-up survey to those values at the end-line survey. For the equating, the Mean & Sigma method is used (Marco 1977).

After the process of equating, the IRT scores of student *i* at the end-line survey, and at the follow-up survey are estimated respectively.

(1) Parameters of the math items at the end-line survey

No.	Item difficulty	Item discrimination	No.	Item difficulty	Item discrimination
1	0.631	0.891	11	1.495	1.401
2	0.510	0.759	12	1.974	1.485
3	2.299	0.292	13	-0.367	1.009
4	-0.150	0.739	14	0.133	2.248
5	0.038	0.756	15	0.489	2.324
6	1.798	1.360	16	3.807	0.893
7	2.774	1.142	17	1.549	1.304
8	3.119	1.160	18	0.845	0.800
9	1.687	1.087	19	2.020	1.270
10	1.057	1.374	20	3.430	0.531

(2) Parameters of the math items at the follow-up survey

No.	Before equating		No.	After equating	
	Item difficulty	Item discrimination		Item difficulty	Item discrimination
1	0.931	0.528	1	1.927	0.454
2	1.713	1.247	2	2.836	1.072
3	0.824	1.197	3	1.802	1.029
4	1.308	1.543	4	2.365	1.326
5	2.676	1.124	5	3.956	0.967
6	1.386	0.911	6	2.456	0.784
7	0.270	1.191	7	1.158	1.024
8	2.109	1.286	8	3.296	1.105
9	2.405	1.142	9	3.641	0.982
10	1.173	0.802	10	2.208	0.690
11	3.767	1.409	11	5.225	1.212
12	1.692	1.028	12	2.811	0.884
13	-1.113	1.196	13	-0.450	1.028
14	1.894	1.263	14	3.046	1.086
15	1.742	0.569	15	2.869	0.490
16	6.878	0.504	16	8.843	0.433
17	0.260	1.674	17	1.146	1.439
18	1.328	1.809	18	2.388	1.556
19	0.583	2.088	19	1.521	1.795
20	0.717	1.741	20	1.678	1.497
21	3.311	1.006	21	4.695	0.865
22	2.297	1.034	22	3.516	0.889
23	4.991	0.624	23	6.648	0.536
24	1.914	0.863	24	3.070	0.742
25	3.160	1.334	25	4.519	1.147

Appendix 8

Principal component analysis of the different types of student household assets

The baseline survey collected information on the following different types of student household assets through student interviews: (a) smartphone; (b) computer; (c) refrigerator; (d) car; (e) television; (f) access to tap water; and (g) flush toilet. In addition, the survey also collected information on the access to electricity and the use of wood for cooking. The data on student household assets consists of dummy variables which take the value of 1 in the case that the student household possess each category of asset that contributes to the wealth of a household. In case of the cooking fuel, as the use of wood is an indicator of the lack of wealth (more wealthy households use gas or electricity for cooking), the non-use of woods in cooking takes the dummy variable representing the asset. A composite index of student household economic status was computed by the principal component analysis.[i] The statistics of first principal component is presented in Table C-1. The first principal component accounts for 25.4 percent of the variances, and all nine asset types have positive loading coefficients, indicating that they contribute to household wealth.

Table C.1. Statistics of first principal component

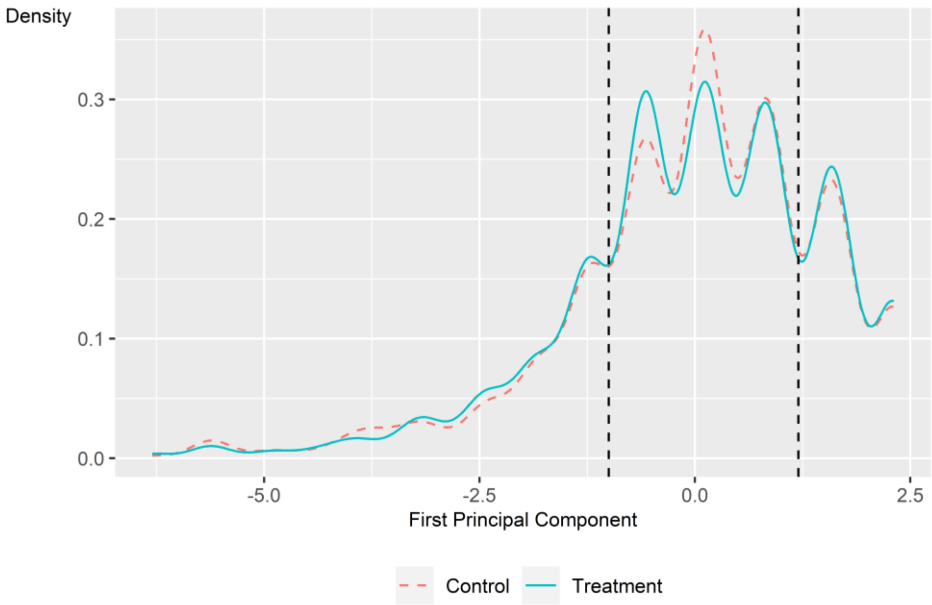
	Coefficient	Correlation with the first principal component
Smartphone	0.238	0.360
Computer	0.390	0.590
Refrigerator	0.402	0.608
Car	0.361	0.546
Television	0.348	0.526
Access to tap water	0.268	0.405
Electricity	0.320	0.484
Flush toilet	0.340	0.513
Non-use of wood in cooking	0.296	0.447

Note: Data Source is the baseline survey of this research.

[i] For Figure 6 in the chapter, the student household economic status index was computed with the baseline sample. For the analysis of heterogeneous impacts in section 3.3.3, the index was computed with the sample of students who were present both at the baseline and end-line surveys. In this appendix, the statistics of the principal component analysis are reported with the sample of students who were present both at the baseline and end-line surveys.

Then, Figure C-1 presents the density curves of the first principal component. Based upon the distribution of the first principal component, the whole sample of students is divided into three groups: lower; medium; and higher household economic status. The dashed lines in Figure C-1 show the threshold values in the first principal component for the groupings. The students with lower household economic status represent around 22 percent respectively in the treatment and control groups. The students with medium household economic status represent around 56 percent in the two groups. The students with higher household economic status represent around 22 percent in the two groups.

Figure C.1: Density curves of the first principal component



Note: Data Source is the baseline survey of this research.

Table C-2 shows characteristics of each household economic status. From lower to higher household economic status, the possession rates of different type of household assets increased. The possession rates of different types of household assets were well balanced between the treatment and the control groups.

Table C.2. Possession rates of different types of household assets by household economic status

Item	Treatment group	Control group	Difference	p value
Lower household economic status (average possession rate %)				
Smartphone	64.6	63.7	0.9	0.844
Computer	2.9	2.6	0.3	0.817
Refrigerator	48.7	44.6	4.1	0.392
Car	2.9	3.9	-1.0	0.428
Television	74.3	75.0	-0.7	0.852
Access to tap water	49.2	54.6	-5.4	0.342
Access to electricity	85.5	83.0	2.5	0.441
Flush toilet	20.7	21.6	-0.9	0.783
Non-use of wood in cooking	12.4	12.6	-0.2	0.926
Number of students	421	388		
Medium household economic status (average possession rate %)				
Smartphone	90.8	89.4	1.4	0.584
Computer	20.3	21.2	-0.9	0.535
Refrigerator	94.8	94.1	0.7	0.516
Car	24	23.1	0.9	0.335
Television	99.2	99.2	0	0.696
Access to tap water	82.8	84.8	-2	0.613
Access to electricity	99.8	99.8	0	0.529
Flush toilet	53.2	57.5	-4.3	0.191
Non-use of wood in cooking	36.5	36.0	0.5	0.896
Number of students	1,038	984		
higher household economic status (average possession rate %)				
Smartphone	98.8	97.9	0.9	0.335
Computer	88.4	85.1	3.3	0.157
Refrigerator	100	100	0	1.000
Car	78.3	80.3	-2.0	0.584
Television	100	100	0	1.000
Access to tap water	97.6	98.4	-0.8	0.492
Access to electricity	100	100	0	1.000
Flush toilet	89.9	94.7	-4.8	0.104
Non-use of wood in cooking	80.4	77.1	3.3	0.443
Number of students	414	375		

Note: Data Source is the baseline survey of this research. The p-values show the results of the chi-squares test with the stratified data (department and urban status).

Table D.1. Comparison of characteristics of schools in four departments (Cabanas, La Union, San Miguel, San Vicente)

Content	Public schools with cycle 3 (A)	Public schools with cycle 1 & 3 (B)	Sampling frame (C)	Sample (D)	Surveyed schools (E)	P-value (D)=(E)
Percentage of N. of schools in urban area	25.7	25.0	29.4	29.2	30.5	0.75
Average N. of students (grade 7) (both shifts)	26.7	26.9	28.6	29.5	30.0	0.88
Average N. of total students (grade 1 to 9) (both shifts)	214.6	216.4	227.7	233.7	237.3	0.88
Percentage of Male students in grade 7	51.6	51.6	51.5	50.7	50.8	0.95
Percentage of grade 7 students in morning shift	52.0	52.0	53.8	55.4	55.7	0.92
Percentage of grade 7 students repeated	4.5	4.4	4.4	4.5	4.6	0.91
Average age of grade 7 students	14.1	14.1	14.2	13.9	13.9	0.96
School infrastructure: Electricity	99.0	99.0	98.9	98.8	98.7	0.96
School infrastructure: Water	82.7	82.5	82.9	83.6	84.1	0.88
School infrastructure: Computer	77.4	77.2	80.7	80.0	80.3	0.93
School infrastructure: Internet	37.8	37.4	40.0	41.2	41.8	0.89
School infrastructure: Library	23.4	23.0	24.2	24.0	24.7	0.86
School infrastructure: Laboratory	9.2	8.6	9.2	9.6	9.6	0.99
School infrastructure: Kitchen	78.7	79.0	78.3	80.0	79.5	0.89
N. of schools	611	605	368	250	239	
N. of schools (which have grade 7)	606	601	366	248	237	

(1) Data source is educational census survey data in El Salvador. Because the data of some schools are not available in the census survey data, the numbers of schools in Column (A) to (C) in this table do not exactly match with those in Table 7. (2) Values on school facilities are binary (Yes:1, No:0). (3) The p-values on number of students, percentage of students in morning shift, and percentage of students who repeated 2nd grade are the results of Wilcoxon rank sum test with stratified data (department dummy, urban/rural dummy). (4) The p-values on binary values show the results of chi-squares test with stratified data (department dummy, urban/rural dummy).

Attrition analysis

Table E.1. Attrition analysis

	OLS End-line (I)-1	Logit End-line (I)-2	OLS Follow-up (II)-1	Logit Follow-up (II)-2
(Intercept)	-0.568^{***}		-0.576^{***}	
	(0.084)		(0.093)	
Treatment	-0.013	-0.013	-0.033	-0.034
	(0.020)	(0.021)	(0.022)	(0.024)
Z score baseline	-0.030^{***}	-0.032^{***}	-0.044^{***}	-0.047^{***}
	(0.006)	(0.007)	(0.007)	(0.008)
Age	0.054^{***}	0.042^{***}	0.068^{***}	0.070^{***}
	(0.006)	(0.005)	(0.007)	(0.008)
Sex (Male=1)	0.022^{**}	0.023^{**}	0.029^{**}	0.031^{**}
	(0.011)	(0.011)	(0.013)	(0.014)
N. of elder brother/sister	0.000	0.000	-0.000	-0.000
	(0.003)	(0.003)	(0.003)	(0.003)
N. of younger brother/sister	0.015^{***}	0.013^{***}	0.009^{*}	0.009^{*}
	(0.005)	(0.004)	(0.005)	(0.005)
Shift (Morning=1)	0.001	0.001	-0.031	-0.032
	(0.026)	(0.024)	(0.031)	(0.032)
Repeated in 2017	0.089^{***}	0.080^{***}	0.076^{***}	0.065^{***}
	(0.020)	(0.019)	(0.021)	(0.022)
N. of asset types	-0.009^{*}	-0.009^{*}	-0.007	-0.008
	(0.005)	(0.005)	(0.006)	(0.006)
La Union	0.073^{**}	0.085^{*}	0.069^{*}	0.075^{*}
	(0.037)	(0.045)	(0.039)	(0.042)
San Miguel	0.098^{***}	0.101^{***}	0.025	0.028
	(0.037)	(0.039)	(0.038)	(0.039)
San Vicente	0.073^{**}	0.077^{**}	-0.004	-0.002
	(0.028)	(0.032)	(0.031)	(0.032)
Urban	0.013	0.006	0.025	0.028
	(0.040)	(0.050)	(0.056)	(0.061)
La Union×Urban	-0.041	-0.029	-0.078	-0.075
	(0.052)	(0.053)	(0.067)	(0.060)
San Miguel×Urban	-0.075	-0.060	-0.021	-0.023
	(0.052)	(0.048)	(0.062)	(0.064)
San Vicente×Urban	-0.049	-0.040	-0.049	-0.057
	(0.048)	(0.050)	(0.057)	(0.058)
Adj. R^2	0.087		0.085	
Num. obs.	4,513	4,513	4,513	4,513
F statistic	19.973		26.376	
N. Clusters	237		237	
Log Likelihood		-2,056.072		-2,572.858
Deviance		4,112.144		5,145.715

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

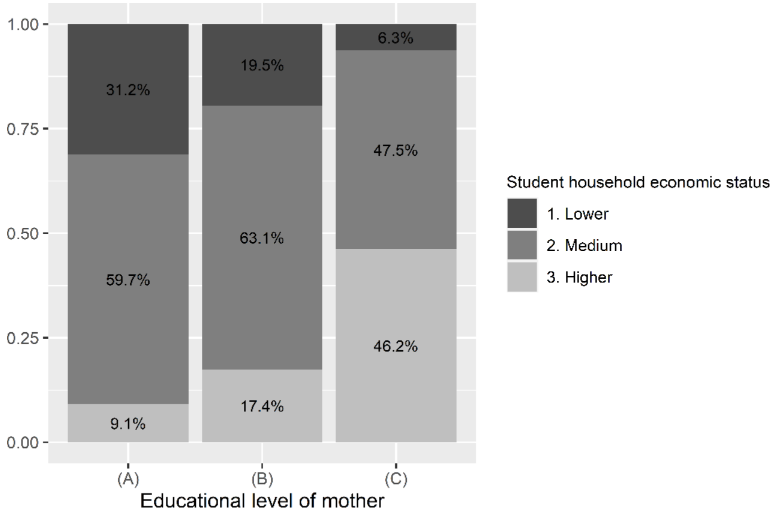
(1) Data source is the baseline survey of this research.

(2) Dependent variable is a dummy that takes 1 for students who were absent respectively at the end- line or follow-up survey. Robust standard errors are clustered at the school level, and are in parenthesis.

(3) Coefficients of the logit regression show marginal effects.

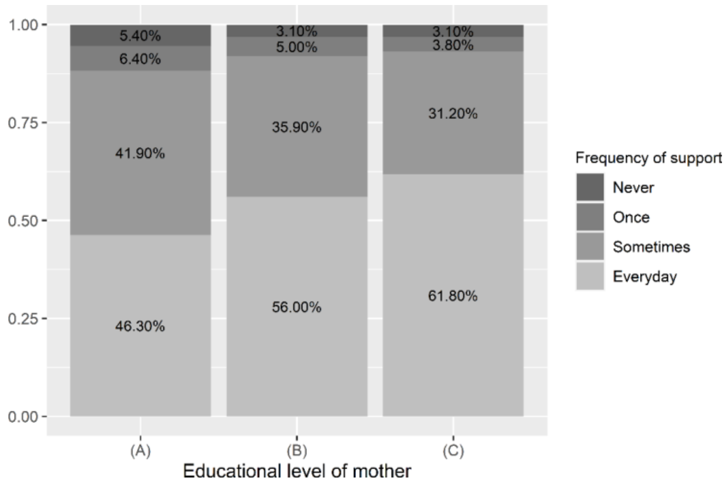
Figures and Table on the analysis of impact of the additional components in the ESMATE program

Figure F.1. Student household economic status by mother's educational record



Note: Data source is the baseline and follow-up surveys of this research. Educational level of mother: (A) never go to school or primary (N.=583); (B) lower secondary (N.=843); (C) upper secondary or university (N.=861). The educational level of mother was surveyed through student interview at the follow-up survey. Among the surveyed students, 826 students answered that they did not know the educational record of mothers.

Figure F.2. Frequency of family support for study at home in a week by mother's educational record

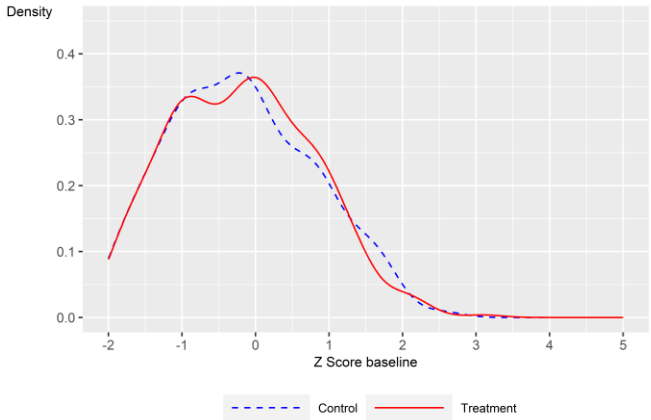


Note: Data source is the baseline and follow-up surveys of this research. Educational level of mother: (A) never go to school or primary (N. =583); (B) lower secondary (N.=843); (C) upper secondary or university (N.=861). The educational level of mother was surveyed through student interview at the follow-up survey.

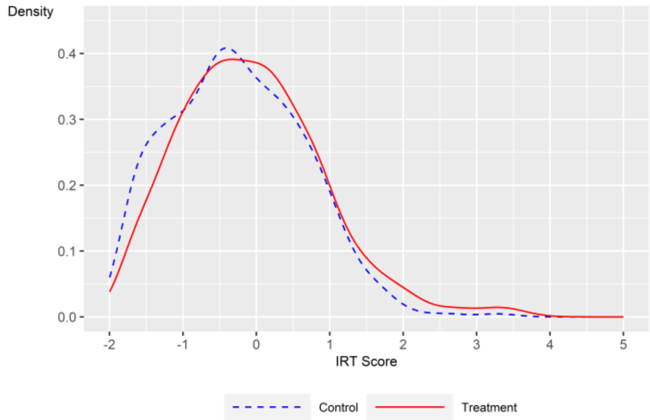
Figure F.3. Density curves of Z scores by the student household economic status levels

(1) Low household economic status

(a) Baseline

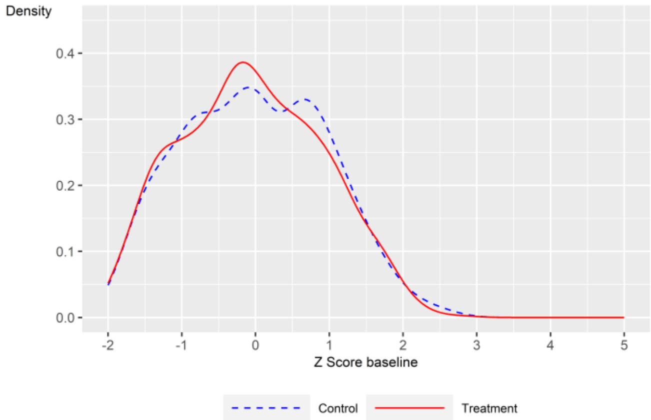


(b) End-line

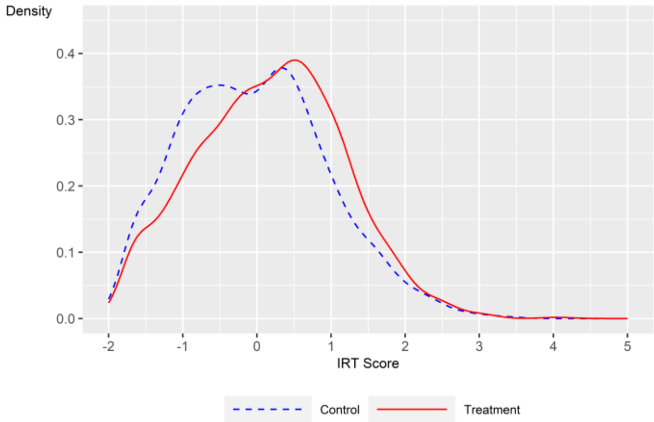


(2) Medium household economic status

(a) Baseline

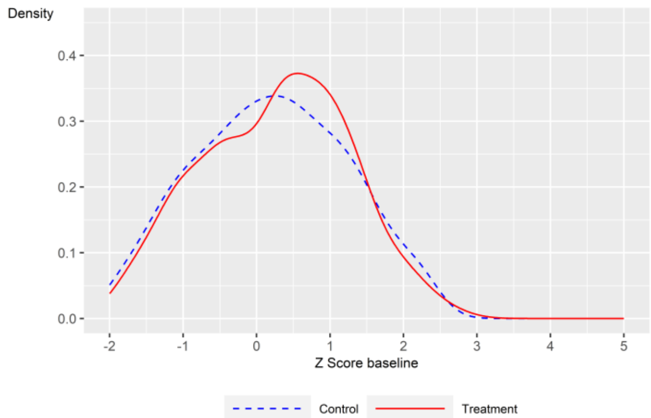


(b) End-line

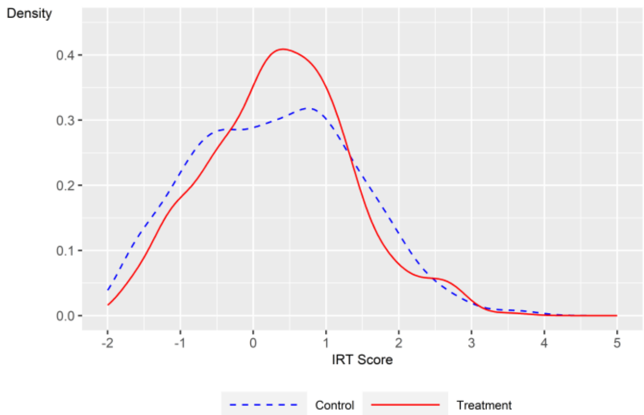


(3) High household economic status

(a) Baseline



(b) End-line



Note: The end-line scores are the IRT scores that are standardized by the mean and standard deviations of the scores of the control group. Baseline test scores are standardized by the mean and standard deviations of the test scores of the control group. Please note that the test scores are not comparable across the baseline and end-line surveys.

Figure F.4. Plot of estimates of the one-year impact by quantile regression

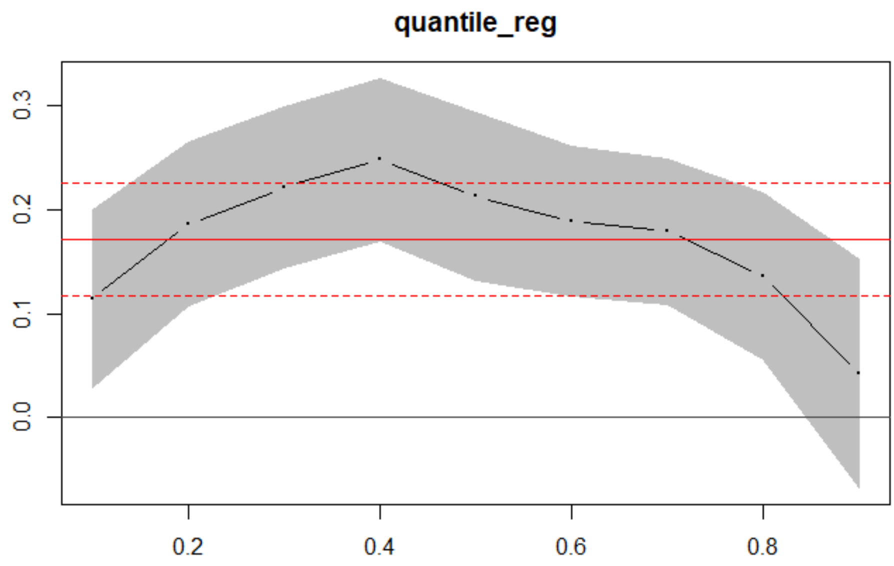


Figure F.5. Plot of estimates of the accumulated impact by quantile regression

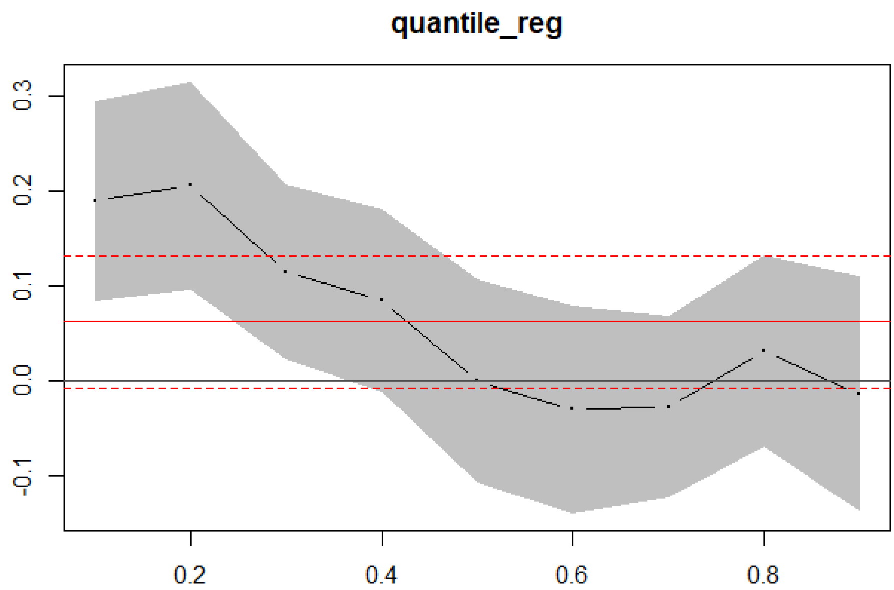


Table F.1. Heterogeneity in the one-year impact of the additional components in the ESMATE program

	(I)	(II)	(III)
(Intercept)	0.006	0.028	0.006
	(0.317)	(0.316)	(0.317)
Treatment	0.210^{***}	0.170^{***}	0.210^{***}
	(0.046)	(0.043)	(0.046)
Treatment × Lower economic status	-0.031		-0.031
	(0.056)		(0.056)
Treatment × Higher economic status	-0.164^{***}		-0.164^{***}
	(0.057)		(0.056)
Treatment × Z score baseline		-0.008	-0.003
		(0.032)	(0.032)
Lower economic status	-0.102^{***}	-0.118^{***}	-0.102^{***}
	(0.038)	(0.030)	(0.038)
Higher economic status	0.122^{***}	0.039^{***}	0.122^{***}
	(0.039)	(0.031)	(0.039)
Z score baseline	0.397^{***}	0.400^{***}	0.398^{***}
	(0.016)	(0.022)	(0.022)
Adj. R^2	0.337	0.335	0.337
Num. obs.	3,619	3,619	3,619
N. Clusters	232	232	232

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$
(1) Data source are the baseline and end-line surveys of this research.
(2) Dependent variables are the IRT scores at the end-line survey. The IRT scores are standardized by the mean and standard deviations of test scores in the control group. Robust standard errors are clustered at the school level, and are in parenthesis.
(3) Baseline test scores are standardized by the mean and standard deviations of test scores in the control group.
(4) Student, school, teacher (end-line) characteristics, and strata fixed effects constructed by department and urban status are controlled but not shown.
(5) In the column (1) to (3), according to the student household economic status index, student samples are divided to lower; medium; and higher household economic status. The student economic status index is the first principal component. Refer to Appendix C for the details.

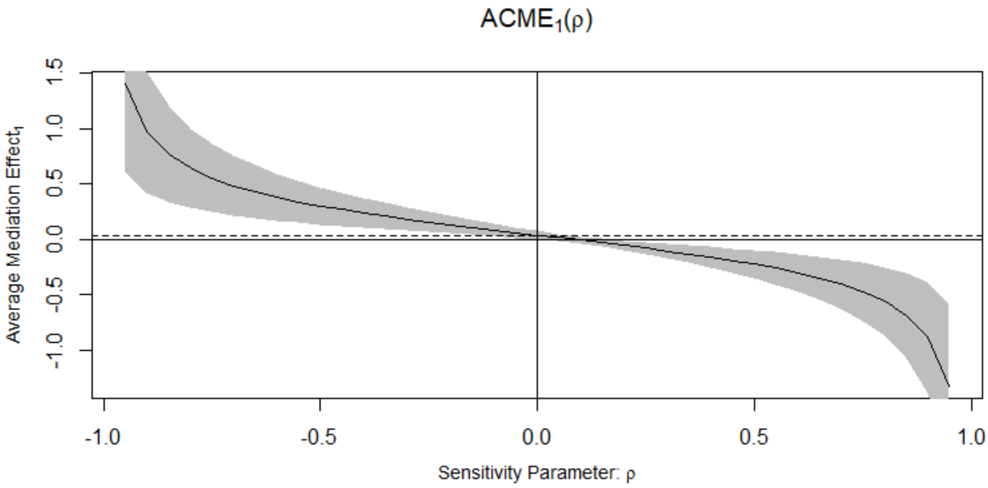
Table F.2. Heterogeneity of the one-year impact of the additional interventions in the ESMATE program on frequency of homework assignment

	(I)	(II)	(III)
Treatment	0.343***	0.337***	0.336***
	(0.088)	(0.087)	(0.087)
Treatment × Lower economic status	0.160*	0.154*	0.151*
	(0.090)	(0.088)	(0.082)
Treatment × Higher economic status	-0.053	-0.040	-0.052
	(0.064)	(0.063)	(0.063)
Treatment × Z score baseline		-0.083***	-0.106***
		(0.030)	(0.033)
Treatment × Lower economic status			0.003
× Z score baseline			(0.058)
Treatment × Higher economic status			0.095
× Z score baseline			(0.042)
Z score baseline	0.016	0.057**	0.060**
	(0.015)	(0.025)	(0.025)
Lower economic status	-0.133*	-0.130*	-0.123*
	(0.074)	(0.072)	(0.066)
Higher economic status	0.042	0.035	0.042
	(0.046)	(0.046)	(0.046)
Lower economic status × Z score baseline			0.038
			(0.052)
Higher economic status × Z score baseline			-0.044
			(0.030)
Num. obs.	3,619	3,619	3,619
N. Clusters	232	232	232

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$
(1) Data source are the baseline, end-line and follow-up surveys of this research.
(2) Robust standard errors are clustered at the school level and are in parenthesis.
(3) In the column (1) to (3), according to the student household economic status index, student samples are divided to lower; medium; and higher household economic status. The student economic status index is the first principal component. Refer to Appendix C for the details.

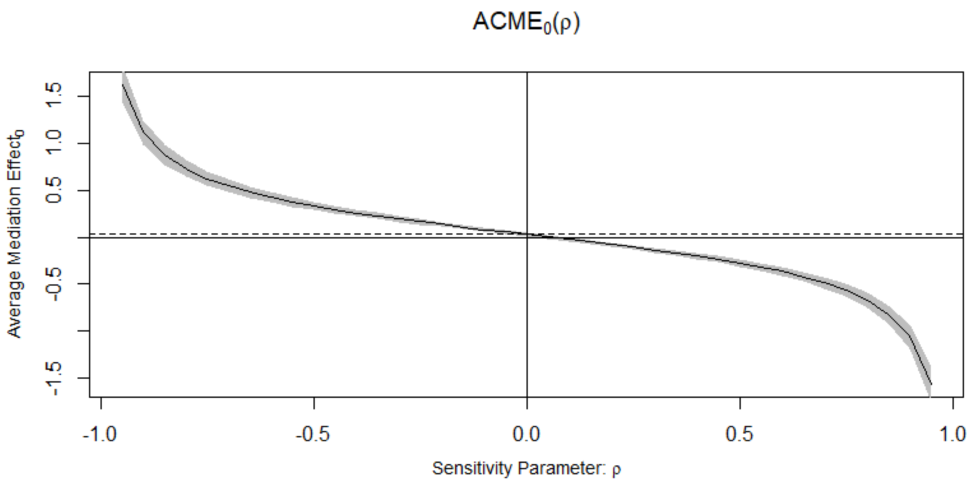
Sensitivity analysis plots

Figure G.1. Plot of sensitivity analysis on the average causal mediation effect of the provision of additional math class for the treatment group



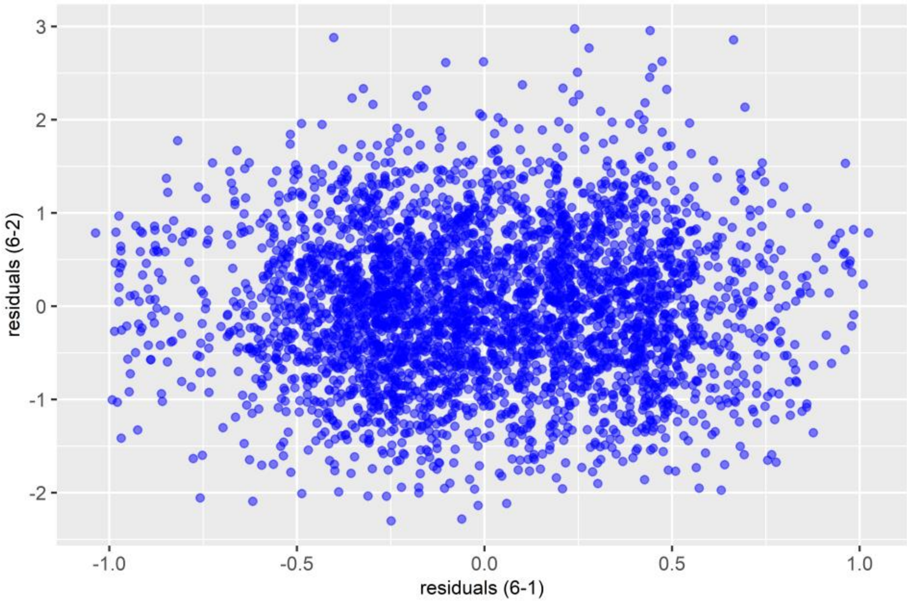
Note: (1) The solid line represents the estimated average causal mediation effect for the given mediator and for different values of ρ, while the shaded area represents the 95 percent confidence interval. The dashed line represents the estimated level of average causal mediation effect when ρ =0. (2) The value of ρ for which the average causal mediation effect becomes zero is 0.10. (3) Data source is the baseline and end-line surveys.

Figure G.2. Plot of the sensitivity analysis on the average causal mediation effect of the frequency of homework assignment for the control group



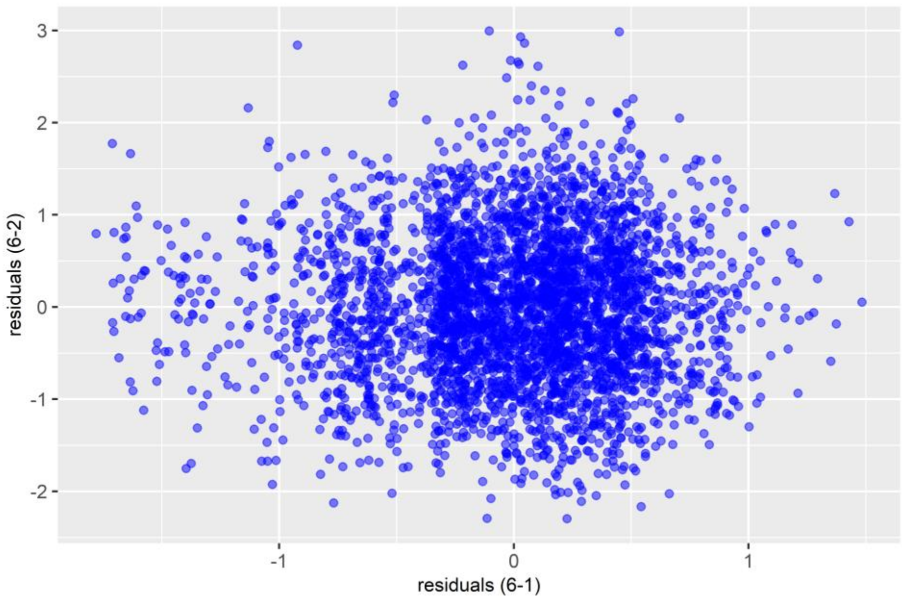
Note: (1) The solid line represents the estimated average causal mediation effect for the given mediator and for different values of ρ, while the shaded area represents the 95 percent confidence interval. The dashed line represents the estimated level of average causal mediation effect when ρ =0. (2) The value of ρ for which the average causal mediation effect becomes zero is 0.05. (3) Data source is the baseline and end-line surveys.

Figure G.3.1. Scatter plot of residuals in (6-1) and residuals in (6-2) in the causal mediation analysis of the provision of additional math class



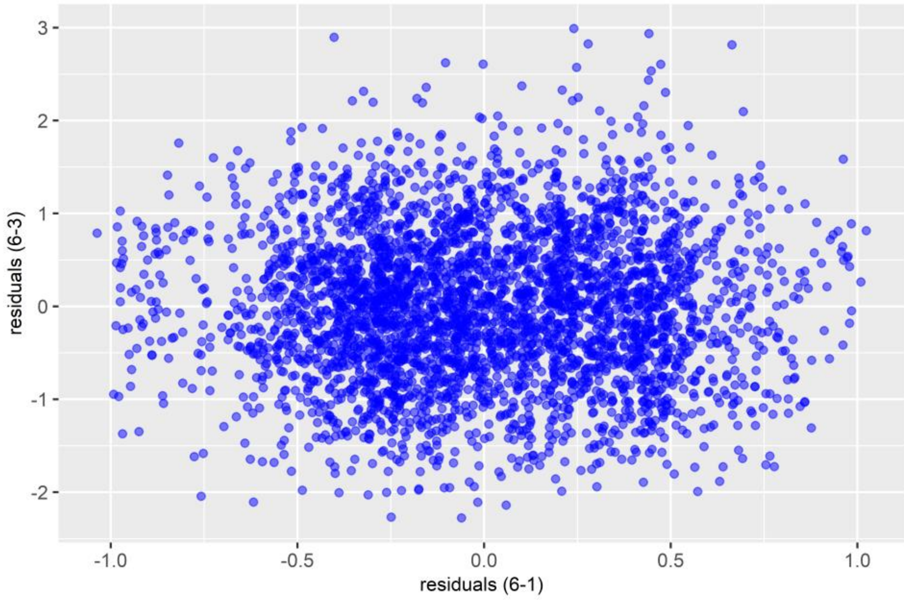
Note: Data source is the baseline and end-line surveys.

Figure G.3.2. Scatter plot of residuals in (6-1) and residuals in (6-2) in the causal mediation analysis of the frequency of homework assignments



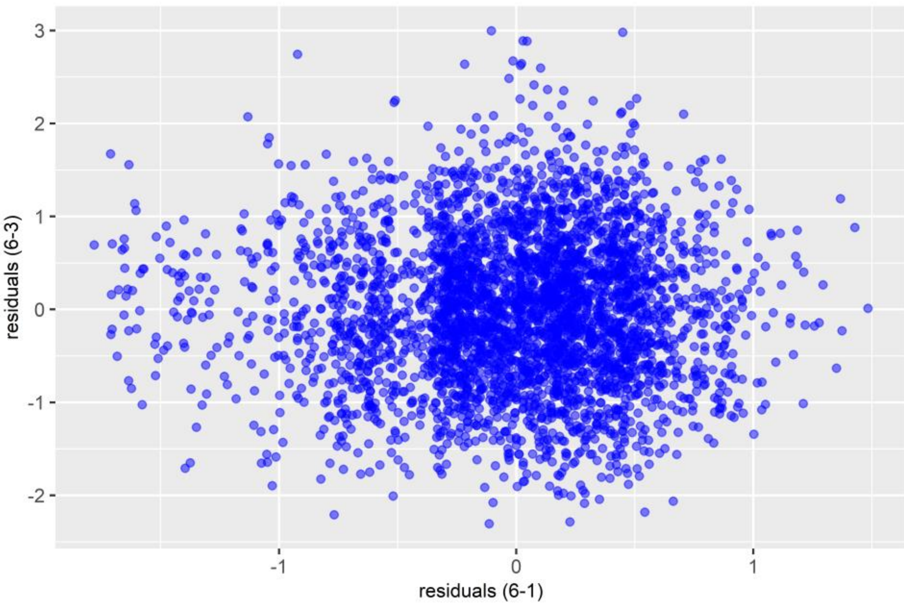
Note: Data source is the baseline and end-line surveys.

Figure G.3.3. Scatter plot of residuals in (6-1) and residuals in (6-3) in the causal mediation analysis of the provision of additional math class



Note: Data source is the baseline and end-line surveys.

Figure G.3.4. Scatter plot of residuals in (6-1) and residuals in (6-3) in the causal mediation analysis of the frequency of homework assignment



Note: Data source is the baseline and end-line surveys.

Chapter 4. Do Remedial Activities Using Math Workbooks Improve Student Learning? Empirical Evidence from Scaled-up Interventions in Niger[45]

4.1 Introduction

In developing countries, curricula expectations are too ambitious, and the mismatch between curricula and children's learning levels expands with each year of schooling (Pritchett and Beatty, 2015). Even though most children are left behind, teachers are required to complete the prescribed syllabus, which means that their attention tends to be selective, often focusing on a limited number of children who can keep up (Banerji, 2000; Abadzi and Llambri, 2011). The simple provision of traditional inputs in educational development like textbooks does not help the learning of children in developing countries who are left behind. Glewwe et al. (2009) revealed that the provision of textbooks written in English only benefits those children with high baseline test scores in Kenya. Since the textbooks were written in English, a third language for most students in the country, the mismatch with student achievement levels hampered most children's learning (Glewwe et al., 2009). The result suggests that the content of learning and the materials should be adjusted to students' learning levels in order to improve their learning. Banerjee et al. (2007) demonstrate that a program specifically targeted to weaker students improved basic reading and mathematics in India. The pedagogical approach employed in the program has been subsequently developed through a series of randomized controlled trials and is now called "Teaching at the Right Level (TaRL)" (Banerjee et al., 2017; Pratham, 2020). Systematic reviews in educational development agree that pedagogical interventions that tailor teaching to student learning levels are effective in improving learning (Evans and Popova, 2016).

It is necessary to scale up successful pilots to address the learning crisis in sub-Saharan Africa; however, the limited capacity of teachers and educational administration can hamper the process. In the 2000s, many countries in the region lowered the level of academic qualifications required to enter teacher training courses (Bashir et al., 2018). The number of primary teachers subsequently increased

[45] This chapter is developed from the following paper and the revised version,
Maruyama, Takao, and Takashi Kurosaki. "Do remedial activities using math workbooks improve student learning? Empirical evidence from scaled-up interventions in Niger." *World Development*, Vol. 148, December 2021, 105659.

by 82.1 percent between 1999 and 2014. Bold et al. (2017) studied primary school teachers' content knowledge in several sub-Saharan African countries and found that the percentage of teachers who met the minimum content knowledge requirements in language and mathematics was low. The rapid expansion of primary enrollment has also outpaced the development of technical capacity in educational administration in the region (Bashir et al., 2018). Local educational administration is too stretched to provide sufficient support for teachers to improve their pedagogical skills. In this context, the distribution of well-structured self-learning materials to students may be an option to support their learning and address the learning crisis. Yet this may pose another challenge. The effective use of such materials depends upon school management and leadership. Sabarwal et al. (2014) found that a large share of the textbooks distributed to schools were being kept in a storage area and not distributed to students in Sierra Leone.

To address the multiple challenges to the improvement of student math learning, Niger has developed a package of interventions: "Paquet Minimum Axé sur la Qualité" (minimum package for quality learning or PMAQ) (Hara et al., 2020). The package was developed by the "Ecole pour tous" project (hereinafter "EPT project") in cooperation with the Japan International Cooperation Agency (JICA). PMAQ seeks to improve math learning in numbers and the four basic operations (addition, subtraction, multiplication and division) through extracurricular remedial activities. The package includes two components. The first is a one-day training of School Management Committee (SMC) presidents and secretaries (school principals), and the second involves the distribution of math workbooks to each student and a two-day training of activity facilitators. Math workbooks are self-learning materials for students. Students receive workbooks that match their learning levels and study math at their own pace. Activity facilitators are teachers and community volunteers who take the role of checking students' answers and giving them instructions. PMAQ was introduced in pilot schools in two of Niger's eight regions (Tillaberi and Niamey). It was found that students who participated in these extra-curricular remedial activities improved their math learning in the 2013-14 school year (JICA 2017).[46]

---

[46] The school year in Niger starts in October and ends in June the next year.

Following this successful pilot, the government of Niger scaled up PMAQ to all public primary schools in the Tillaberi region in the 2017-18 school year. The scaled-up interventions were aimed at helping the students improve their math learning through extra-curricular remedial activities over a period of approximately three months. However, because of budget constraints, the distribution of math workbooks was limited to the 1st to 4th grades. Focusing on the difference in the intervention between 4th and 5th grade students, this study employed the difference in differences (DD) to investigate the impact of PMAQ on student math learning. The average impact of the package on student math learning is estimated to be 0.36 to 0.38 standard deviations of math test scores.

This chapter investigates the impact of scaled-up interventions on student math learning using quasi-experimental data from three rounds of survey. While the number of studies looking at pilot interventions has increased, rigorous evidence on scaled-up interventions is still scarce in the education field. Given the challenges of implementation, the success of a small pilot does not guarantee the success of a scaled-up intervention (Banerjee et al., 2017; Bold et al. 2018). Scaling-up is often implemented with reduced costs from the pilot model to cover more schools. It could result in attenuating the impact of the original model because of the reduction in the provision of learning materials and direct support for teachers, and the complementarities among the different inputs (Kerwin and Thornton, 2020).

This chapter evaluates the heterogeneity of impacts with respect to two aspects that are important in Niger. One of them is the baseline score level. In Niger, 92.4 percent of students in the last grade of primary education do not reach sufficient levels in mathematics, and 68.4 percent cannot solve basic math problems such as the four basic operations (PASEC, 2016). The achievement level for girls in mathematics is lower than that of boys (PASEC, 2016). Thus, this study also assesses the heterogeneous impact with respect to gender. It turns out that while the impact is larger on students with lower baseline scores, the impact additional for girls is not statistically significant.

The results of the analysis in this chapter demonstrate that a package of interventions can address multiple challenges including the limited capacity of teachers and educational administrations to improve student math learning. In Niger, district educational offices suffer from a lack of staff and logistical resources. In 2018, each district educational office oversaw on average 249 School

Management Committees (SMCs) (Kunieda et al., 2020). Less than one percent of primary school teachers possess minimum knowledge of the subjects that they teach (World Bank, 2017). Niger is one of the countries most affected by the learning crisis. *Le Programme d'analyse des systèmes éducatifs de la CONFEMEN* (PASEC), the regional standardized exam mainly for French-speaking sub-Saharan African countries, shows that among the eleven countries surveyed, the percentage of students in the last grade of primary education that were above the minimum competency threshold was lowest in Niger (PASEC, 2016). Bashir et al. (2018) categorized the countries of sub-Saharan Africa into the following four groups according to educational performance in terms of access and quality: (1) Established; (2) Emerged; (3) Emerging; and (4) Delayed; Niger is included in the last group. If a package of interventions addresses the multiple challenges to improving student learning in Niger, the package likely helps the other countries in the region to overcome the learning crisis.

The remainder of this chapter is organized into the following five sections. Section 4.2 describes the content of the interventions and the timeline for the scaling up. Section 4.3 explains the evaluation strategy to estimate the impact of scaled-up interventions on student math learning. Section 4.4 reports the results, and section 4.5 then discusses the impact on student math learning and the heterogeneity. Section 4.6 concludes.

4.2 Content of interventions and the timeline for the scaling up

4.2.1 Content of interventions

PMAQ seeks to improve math learning through extracurricular remedial activities organized by SMC.[47] For this purpose, the package was composed of two components. The first is a one-day training course for SMC presidents and secretaries (school principals) that strengthens the capacity of SMCs to develop a school action plan. The action plan integrates extra-curricular remedial activities that include the use of math workbooks.[48] After the training, each school conducts a test on the four

47 For further details about PMAQ, see Hara et al. (2020) and Kunieda et al. (2020).

48 In Niger, the Ministry of Education organized a set of trainings on the democratic establishment of school management committees and school action plan development cycles for all public primary schools in the 2007-08 and the 2008-09 school years (Honda and Kato, 2013). The school action introduced at the time did not address the quality of education. The first component of PMAQ tried to improve the functionality of school management committees to develop school action plans that address improvements in learning.

basic operations in math to measure students' learning levels. Then, SMC organizes a community general assembly to share the results among teachers, parents, and community members, and to discuss the causes of low levels of learning and potential activities that may improve the situation.[49] Based upon these discussions, SMC prepares a draft of a school action plan that integrates remedial activities. The plan is adopted at the following community general assembly. Parents commit themselves to send their children for extra-curricular remedial activities (Hara et al., 2020).

The second component of PMAQ is pedagogical and consists of the distribution of math workbooks accompanied by a two-day training course for activity facilitators (teachers and community facilitators). The EPT project has developed nine workbooks on numbers and the four basic operations, which reflect the different math learning levels.[50] Based on the results of the math assessment, each primary student is given math workbook that matches their learning level. Students learn math by using the workbooks at their own pace. Since students write in the workbook when they study, the workbooks cannot be reused for other students. For the same reason, the workbook cannot be shared with other students. Activity facilitators check the answers and give instructions to each student. The content of the workbook is systematically arranged so that students can easily follow and progressively understand more mathematics principles. For example, students learn what numbers are by following the workbook pages that gradually transform an illustration of a concrete object into a number. The math workbook

---

49 SMC facilitates the discussion at the community general assembly using the analysis framework: (i) inadequate time spent on learning; (ii) lack of suitable learning materials and environments; and (iii) ineffective and low-quality teaching and learning practices. For each of these factors, the community identifies a corresponding solution (e.g. increasing the amount of time spent on learning, providing suitable learning materials and environment, and improving the quality of teaching and learning practices). These solutions are then integrated into the SMC-organized extracurricular remedial activities.

50 The math workbooks were issued in nine volumes:
Volume 0 (pre-math): Introduction to handwriting.
Volume 1: Introduction to numbers: 0 – 10; Composition and decomposition of numbers.
Volume 2: Introduction to numbers: 11-20; Addition/subtraction without carrying and borrowing; Addition and subtraction with carrying and borrowing.
Volume 3: Addition and subtraction of 3-digit numbers; Understanding numbers: 21-79.
Volume 4: Addition and subtraction of 2-digit numbers; Number sense 80 – 120; Addition and subtraction: vertical; Number sense: up to 1,000.
Volume 5: Addition and subtraction of 2 to 3-digit numbers; Number sense: up to 10,000; Addition and subtraction of 2 to 3-digit numbers: vertical.
Volume 6: Multiplication and division: basic; Division with remainders.
Volume 7: Number sense: 10,000 – 100,000,000; Multiplication: vertical.
Volume 8: Division: vertical; Calculation rules and techniques.

includes plenty of well-developed math problems so that students can practice mathematics skills until they have mastered them.

PMAQ was introduced in pilot schools in the Tillaberi and Niamey regions of Niger's eight regions. The Tillaberi region is situated in the southwest of the country, surrounding the capital city, Niamey. The population of the Tillaberi region is around 2.6 million, accounting for around 16 percent of the country's population. Of the population, 53.8 percent live in poverty (national average: 59.5%). While access to primary education in the region is close to the national average, the quality of education is lagging behind. Of the students in the last year of primary school in the region, 97.1 percent do not reach sufficient levels in mathematics (PASEC, 2016). The gender gap is also an issue in educational development. The primary gross enrollment ratio in the region is 79.3 percent (boys) and 74.2 percent (girls), and the primary completion rate is 75.3 percent (boys) and 67.2 percent (girls) (The Ministry of Primary Education in Niger, 2016). In such a difficult context, students who participated in extra-curricular remedial activities in the pilot schools in the 2013-14 school year showed an improvement in their math learning (JICA, 2017). As a result, the Government of Niger decided to scale up PMAQ in the Tillaberi region in collaboration with multi-donor support.[51]

Remedial activities were organized from March to June 2018 for the students in the Tillaberi region; however, because of budget constraints, the distribution of math workbooks was limited to students in the 1st to 4th grades of the six grades in primary education.

4.2.2 Timeline for the scaling up of interventions

The Ministry of Education in Niger initially scheduled the scaling up of PMAQ for the 2016-17 school year. From November to December 2016, the training of SMC presidents and secretaries (school principals) on school action plan development was organized. Math workbooks were to have been delivered by March 2017. However, as the procurement process was significantly delayed, the

[51] Multi-donor support was provided by the Global Partnership for Education, AFD, Swiss Cooperation, and the World Bank. In the 2013-14 school year, PMAQ was piloted in Karma commune and a part of Niamey city. The communes were not selected in the process of sampling schools for the survey in 2017 and 2018. Thus, there is no duplication between the pilot schools in the 2013-14 and the schools surveyed in 2017 and 2018 that this chapter discusses.

Ministry postponed both the distribution of math workbooks and the training of activity facilitators to the following school year.

In February 2018, a two-day training on the use of math workbooks was organized for activity facilitators who were selected by each SMC. In March 2018, before the start of the remedial activities at the SMCs, the Ministry organized a forum of stakeholders including departments of the Ministry, the director of the local education administration office, inspectors, pedagogical advisors, executive members of the SMC federation, SMC supervisors, and mayors (The EPT project, 2018).[52] At the forum, the target for improvements in math learning was discussed, and the participants committed to a 30 percentage points increase in the correct response rate of primary 1st to 4th grade students through remedial activities. They also pledged to organize at least 10 hours of remedial activities per week at each SMC.[53]

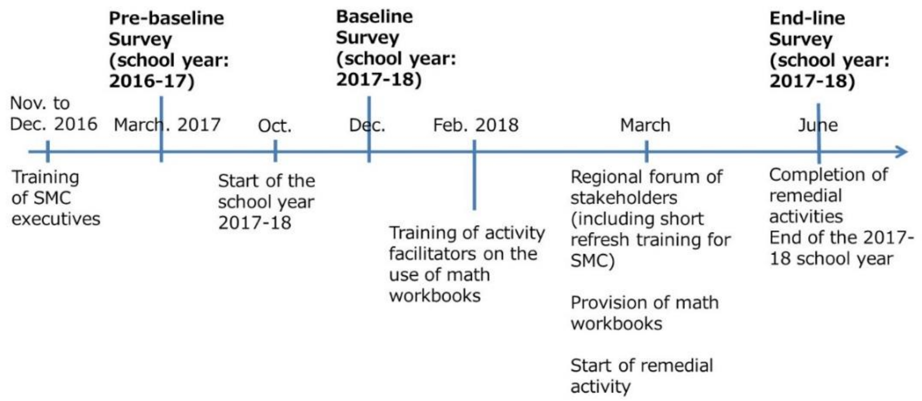


Figure 10. Timeline for the scaling up of interventions

According to the Ministry of Primary Education in Niger (2018), 3,380 targeted primary schools planned and conducted remedial activities in math. Each SMC implemented an average of 133 hours of extracurricular remedial activities in math (around 10 hours per week) in the three months between March and June 2018. These activities were supported by 12,735 facilitators, of whom 86 percent were teachers. Once these remedial activities were completed, the correct responses given by

52 The SMC federation is a network of SMCs in a geographical or administrative area (municipality or district) that supports SMCs in its network.
53 Each SMC federation organized a general assembly to report and discuss the results of the forums with member SMCs, and to pledge to achieve the targets. Subsequently, each SMC organized a community general assembly to discuss the objectives, planning, and implementation arrangements for remedial activities.

students during basic math assessments significantly improved from a starting point (baseline) of around 40 percent to 67 percent on average in the end-line (The Ministry of Primary Education in Niger, 2018). However, these figures are highly inadequate as evidence for the impact of scaling up the interventions. First, this is a before and after comparison and does not control for changes over time that were not related to the project. Furthermore, the data quality is not satisfactory.[54] Therefore, this chapter estimates the impact in a more rigorous way, as explained in the next section.

4.3 Evaluation strategy

4.3.1 Sampling of schools and student test data

Of the 3,655 public primary schools in the Tillaberi region, 3,519 schools in 13 departments taught at least one class in the 1st to 4th grades. Niger has two types of public primary schools: those that teach in French (including those that teach in French plus the local language); and (ii) those that teach Arabic in addition to other subjects. The latter, known as "Franco-Arab primary schools," constitute around 10 percent of the public primary schools in the region.

The EPT project selected 31 public primary schools that teach subjects in French and 3 Franco-Arab primary schools. Among the former, 28 schools were selected through three-stage random sampling: at the departmental, commune, and school levels.[55] The remaining three were schools where the EPT project piloted an upgraded version of PMAQ that integrates basic reading and writing in addition to the math components.[56] The three Franco-Arab primary schools were randomly chosen

---

[54] The SMC conducted basic math assessments and reported the results to the local education office, which then compiled and forwarded them to the Ministry of Primary Education. There are several potential biases in the student assessment results reported by each SMC and compiled by the Ministry of Primary Education. First, as the math test was voluntarily organized by each SMC, the quality of administrative control over the conducting of the tests varied among SMCs. Second, since it was not mandatory for SMCs to submit activity reports to the local education administration office, not all SMCs reported the test scores. Third, the numbers in the SMC activity reports were summed up at the local education administration offices. Due to the large volume of assessment results, it is possible that in the summing up process, the numbers may not have been correctly calculated. In addition to those potential biases, the student-level individual test results were not available in the data compiled by the Ministry. Given those limits, this study uses the data collected by a local NGO under contract with the EPT project.

[55] Five departments situated mainly in the northern part of the region were excluded from the sampling frame because of security concerns.

[56] For the three schools, math workbooks were delivered by the project. On the other hand, other schools came to the local education office to receive workbooks. The transportation expenses were borne by the multilateral support donors.

from the same departments selected in the above-mentioned random sampling process. As a result, the evaluation sample is made up of 34 schools, consisting of 31 public primary schools and 3 Franco-Arab primary schools (Table 13).

A local NGO conducted three rounds of surveys: in March 2017 (pre-baseline[57]); December 2017 (baseline); and June 2018 (end-line).[58] As explained in the previous section, due to the delay in the procurement of math workbooks, the planned scaling up of PMAQ was postponed. The first survey in March 2017, which was initially designed as the baseline survey, became the pre-baseline survey. As the school calendar in Niger starts in October and finishes in June the next year, the pre-baseline survey was conducted in the 2016-17 school year, and the baseline and end-line surveys were organized in the 2017-18 school year.

Table 13. Distribution of sampled schools by department and commune

Types of public primary school	Department	Commune	N. of schools which have both 4th & 5th grades (t1 and t2) (A)	N. of schools which have only 4th grade (t1 and t2) (B)	N. of schools which have only 5th grade (t1 and t2) (C)	N. of schools in the evaluation sample of this study (t1 and t2) (A)+(B)+(C)
Public primary school which teaches subjects in French or local language	Kollo	Kollo	2	4	0	6
		Hamdalaye	2	1	1	4
	Say	Say	3	1	0	4
		Gueladio	0	1	3	4
	Gotheye	Gotheye	1	2	0	3
		Dargol	2	1	0	3
	Torodi	Makalondi	2	1	1	4
		Torodi	3	0	0	3
Franco-Arab	Kollo	Kollo	1	0	0	1
		N'Dounga	0	0	1	1
	Torodi	Torodi	0	1	0	1
Total			16	12	6	34

Note t1: baseline survey; t2: end-line survey.

57 The remedial activities with math workbooks were implemented from March to June 2018. The survey in December 2017 describes situations just before these interventions. As this survey provides the most appropriate reference information to evaluate their impact, this study calls the December 2017 survey the "baseline survey". Accordingly, this study call the survey in March 2017 the "pre-baseline" survey.

58 The authors were not involved in the survey design and data collection process. School sampling, preparation of test items, and surveys were conducted by the EPT project and a local NGO. JICA provided all data used in this study.

In each round of surveys, teams from a local NGO organized math tests to measure the learning outcomes of students in all grades at each school.[59] For each grade, the test was composed as follows:

- Grade 1: section (i) single-digit addition (12 questions);
- Grade 2: section (i) plus section (ii) single-digit subtraction (12 questions);
- Grade 3: sections (i) and (ii) plus section (iii) two-digit addition including problems with carrying, and subtraction including problems with borrowing (11 questions) and (iv) single-digit multiplication (9 questions);
- Grade 4: sections (i) through (iv) plus section (v) three-to-four-digit addition including problems with carrying and subtraction including problems with borrowing (12 questions); and
- Grades 5 and 6: sections (i) through (v) plus section (vi) single-digit division including problems with remainders (5 questions).

There were 44 test items for 3rd grade students, 56 for 4th grade students, and 61 for 5th and 6th grade students. To compare the test scores of 4th and 5th grade students, 56 items for 4th grade students are used.

The characteristics of 56 math items in relation to their difficulty and discrimination are checked by the baseline data of 4th grade students. The results are shown in Appendix 13 of this chapter. There is a variety of correct response rates among the 56 items. In terms of the item-total scores correlations, none of the values are close to zero. When the correlation value of item *n* is close to zero, this indicates that the item does not function well in measuring student math achievement levels. We also calculated Cronbach's alpha for each section of the math test as the reliability measure. All of the values are larger than 0.8, indicating that the internal consistency of test items by each section is sufficiently high. A two-parameter logistic model of the item response theory is used to estimate the parameters of item difficulty and item discrimination. The estimated parameters of item difficulty vary among the items, and those relating to item discrimination are not low. The results indicate that the

[59] The test took 30 minutes for the 1st and 2nd grades and 45 minutes for the other grades.

items of the math test were able to capture the different levels of student math achievement on the four arithmetic operations well.

4.3.2 Constructing panel data

Since the three survey rounds were not designed to track the same students, this study constructed panel data on 4th through 6th grade students ex-post by manually linking the data from the three rounds of survey. As student names were typed at each round of the survey, there was potential for inconsistencies in the spelling of names between the three rounds. To overcome such inconsistencies in the spelling of student names, this study utilized a fuzzy-joining method (Robinson, 2019). Personal names in Niger generally consist of an individual's first name, father's first name, and grandfather's first name. This study fuzzy-joined three waves of survey data using the letters of the student's first name, the letters of the father's name, and the name of the school. The process of constructing panel data is described in Appendix 14. The panel data of two data points (baseline and end-line) includes 1,070 students from the 4th through 6th grade, and that of all three data points contains 717 students. The gender of the students was not collected in any of the three survey rounds. This study recovered this information based on the typical correspondence between first names and gender. There were approximately 700 different first names in the pooled data from the baseline and end-line surveys, which this study successfully classified into either boy's or girl's names.[60]

Table 14 presents the descriptive statistics of the math test scores in each round of the surveys by different data types. The baseline score level of the 4th grade students is lower than that of the 5th graders. The average scores and the standard deviations of 4th and 5th grade students are similar across the different types of data. Then, Graph 1 shows the trends in the average math test scores of the 4th and 5th grade student group in the three survey rounds. The "4th grade student group" refers to those who were 4th grade students during the 2017-18 school year, including students who had moved up from the 3rd grade or repeated the 4th grade. The "5th grade student group" refers to those who were in the 5th grade during the 2017-18 school year and includes students who had moved up from the 4th

[60] The EPT project staff provided information on the association between first name and gender. In Niger, boys' and girls' names are rarely gender-ambiguous.

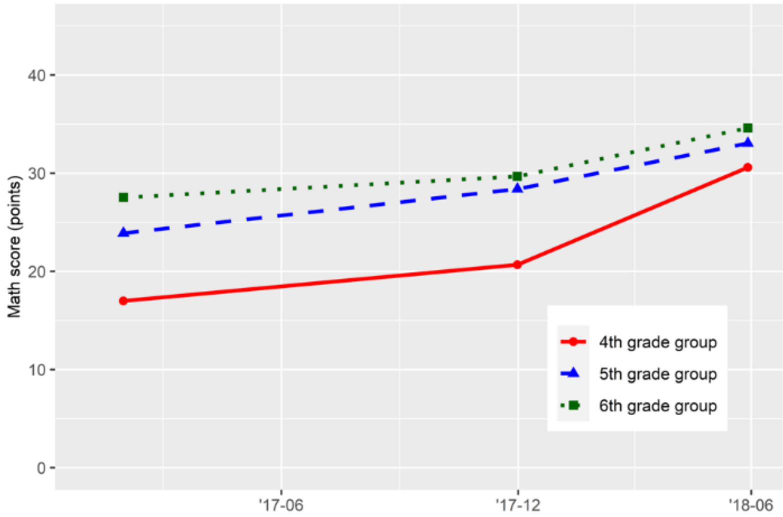
grade or repeated the 5th grade. The trends in the average math test scores of the 4th and 5th grade student groups are parallel from the pre-baseline to baseline surveys. However, the 4th grade average math test scores improved more than the 5th grade student group between the baseline and end-line surveys. This outcome is analyzed in the next section.

Table 14. Number of students and math test scores by grade group

Data type	Pooled data (2t)			Panel data (2t)			Panel data (3t)		
Maximum scale of test scores	Maximum scale (full mark) of test scores: 56			Maximum scale (full mark) of test scores: 56			Maximum scale (full mark) of test scores: 44		
Grade	Grade 4	Grade 5	Grade 6	Grade 4	Grade 5	Grade 6	Grade 4 group	Grade 5 group	Grade 6 group
Pre-baseline									
Average test scores							17.1	23.9	27.5
Standard deviation							13.0	12.9	12.5
N. of students (girls)							120	98	129
N. of students (boys)							116	119	135
Baseline									
Average test scores	22.0	31.0	35.1	23.5	32.4	35.0	20.7	28.4	29.7
Standard deviation	15.4	14.7	14.6	15.6	14.3	14.3	13.6	12.0	11.3
N. of students (girls)	288	237	276	193	148	196	120	98	129
N. of students (boys)	257	284	295	164	181	188	116	119	135
End-line									
Average test scores	34.3	37.7	40.6	35.7	38.5	41.0	30.6	33.0	34.6
Standard deviation	14.7	14.2	14.1	14.2	13.7	14.0	11.7	10.4	10.4
N. of students (girls)	290	199	272	193	148	196	120	98	129
N. of students (boys)	226	214	264	164	181	188	116	119	135

Notes:
(1) "Grade 4 group" is defined as a group of students who were in grade 4 during the 2017-18 school year. "Grade 5 group" and "Grade 6 group" are defined in the same way. (2) The first section "Pooled Data" refers to the test scores of students in the 34 schools reported in Table 13, which includes students who were not linked across different survey rounds. The other two sections are subsets of the first section. The second section "Panel Data (2t)" refers to 1,070 students who were successfully linked between the baseline and end-line surveys. The third sections cover 717 students who were successfully linked between the pre-baseline, baseline, and end-line surveys. (3) For the comparison of test scores of the panel data (3t), 44 items (only the items from sections (i) through (iv) corresponding to the items in math test for grade 3) are used.

Graph 1. Trends in average math test scores (4th to 6th grade groups)



Notes: (1) The "4th grade group" is defined as a group of students who were in the 4th grade during the 2017-18 school year. The "5th grade group" and "6th grade group" are similarly defined.
(2) The number of students in the 4th grade group is 236; the number of students in the 5th grade group is 217; and the number of students in the 6th grade group is 264.

4.3.3 Estimation Strategy

To estimate the impact of PMAQ on student learning outcomes, this study uses a difference in differences (DD) method. As noted, because of budget constraints, the distribution of math workbooks was limited to students from the 1st to 4th grades. While 4th grade students were targeted by remedial activities that utilized math workbooks, 5th grade students were not. Focusing on the difference in the intervention between 4th and 5th grade students, this study first estimates a DD specification without student fixed effects using unbalanced pooled data. The DD specification is depicted as:

$$Y_{ijt} = \alpha + \beta_1 \, Grade4_i + \beta_2 \, Post_t + \lambda \, (Grade4_i \times Post_t) + X_i\gamma_1 + X_j\gamma_2 + \varepsilon_{ijt} \, , \quad (1)$$

Y_{ijt} is the standardized test scores of student i in school j at the time of the tth survey; $Grade4_i$ is a dummy variable that takes the value 1 for the student i of the 4th grade in the 2017-18 school year, and takes 0 for the 5th grade; $Post_t$ is a dummy variable that takes the value 1 for the end-line data; X_i is a vector of student characteristics for student i, including the female dummy; X_j is a vector of school characteristics for school j, such as the dummy for a Franco-Arab school or a pilot school of upgraded version of PMAQ that integrates basic reading and writing activity in addition to math components; ε_{ijt} is an error term.

Math test scores are standardized by the mean and the standard deviations of the baseline scores of 4th grade students. The coefficient of interest is λ, which identifies the average impact of PMAQ on 4th grade students. Scalers of parameters α, β_1, β_2 and λ, and vectors of parameters γ_1 and γ_2 are to be estimated econometrically. The robust standard errors of parameters are clustered at the school level.

As the main focus of this study is not on the impact of school characteristics on math learning, $X_j\gamma_2$ in equation (1) is replaced with school fixed effects, α_j. Since the treatment assignment was done by different grades in the same schools, it is possible to estimate the impact of PMAQ more precisely by controlling the school fixed effects. Equation (1) then becomes:

$$Y_{ijt} = \alpha_j + \beta_1 \text{Grade4}_i + \beta_2 \text{Post}_t + \lambda (\text{Grade4}_i \times \text{Post}_t) + X_i\gamma_1 + \varepsilon_{ijt}, \quad \text{(1-a)}.$$

By using the balanced panel data of two periods instead of the unbalanced pooled data, it is possible to control the student characteristics more flexibly, using the student fixed effects α_i. The estimation model then becomes:

$$Y_{ijt} = \alpha_i + \beta_2 \text{Post}_t + \lambda(\text{Grade4}_i \times \text{Post}_t) + \varepsilon_{ijt}, \quad \text{(2)}.$$

By controlling for the student fixed effects, it is possible to estimate the average impact of PMAQ more precisely. While there would be differences in observed and unobserved characteristics between 4th and 5th grade students, the DD specification with the student fixed effects cancels out the effect of time-invariant characteristics of each student on math learning outcomes. Note that each student attends the same school in both the baseline and end-line surveys. In other words, student fixed effects and school fixed effects are perfectly collinear, rendering the latter redundant. By construction, parameters β_1, γ_1, and γ_2 in equation (1) cannot be identified in specification (2).

The assumption to identify λ is the parallel trends in the test scores of 4th and 5th grade student groups. If the parallel trends assumption does not hold, λ cannot correctly identify the impact of PMAQ. The parallel trends assumption is checked by equation (3) with the pre-baseline and baseline data. *Pre-Treatment$_t$* is a dummy variable that takes the value 1 for the baseline data. If the parallel trend assumption holds, λ' will be close to zero and will not be statistically significant.

$$Y_{ijt} = \alpha_i + \beta_2 \text{ Pre-Treatment}_t + \lambda' (\text{Grade4}_i \times \text{Pre-Treatment}_t) + \varepsilon_{ijt}, \quad (3)$$

As another test for the parallel trends assumption, this study also check the trends in math learning among students in different grades who were not targeted through remedial activities involving the use of math workbooks. More concretely, this study estimate equations (4) and (5) by comparing 5th and 6th grade students in the 2017-18 school year. *Grade5$_i$* is a dummy variable which takes the value 1 for 5th grade students in the 2017-18 school year, and 0 for the 6th grade. If the trends in test scores of those grade students are parallel, λ'' will be close to zero and not be statistically significant.

$$Y_{ijt} = \alpha_j + \beta_1 \text{ Grade5}_i + \beta_2 \text{ Post}_t + \lambda''(\text{Grade5}_i \times \text{Pre-Treatment}_t) + X_i\gamma_1 + \varepsilon_{ijt}, \quad (4)$$

$$Y_{ijt} = \alpha_i + \beta_2 \text{ Post}_t + \lambda''(\text{Grade5}_i \times \text{Pre-Treatment}_t) + \varepsilon_{ijt}, \quad (5)$$

4.4 Results

4.4.1 Average impact on math learning

The average impact of PMAQ on student math learning is estimated at 0.36 standard deviations by equation (1) with the pooled data (Table 15 (I)), and at 0.38 standard deviations by equation (1-a), which controls school fixed effects (Table 15 (II)). The estimate remains at the same level in the regression analysis by equation (1-a) and (2) using the panel data (Table 15 (III) and (IV)). All estimates are statistically significant at the 1 percent level.[61]

Table 15. Average impact of PMAQ on math learning

	(I) **4th / 5th** Pooled Model (1)	**(II)** **4th / 5th** Pooled Model (1-a)	**(III)** **4th / 5th** Panel: 2t Model (1-a)	**(IV)** **4th / 5th** Panel: 2t Model (2)
(Intercept)	0.702***			
	(0.159)			
Grade 4 (Treatment)× Post (λ)	0.367***	0.387***	0.387***	0.387***
	(0.088)	(0.081)	(0.076)	(0.107)
Grade 4 (Treatment)	-0.540***	-0.625***	-0.511***	
	(0.137)	(0.102)	(0.109)	
Post	0.435***	0.422***	0.394***	0.394***
	(0.065)	(0.067)	(0.055)	(0.076)
Female	-0.207**	-0.213***	-0.160**	
	(0.076)	(0.058)	(0.074)	
Franco-Arab school	-0.197			
	(0.217)			
Basic reading pilot school	-0.333			
	(0.241)			
School fixed effect (dummy)	No	Yes	Yes	No
Student fixed effect (dummy)	No	No	No	Yes
Adjusted R-Squared	0.151	0.508	0.511	0.841
Observations	1,995	1,995	1,372	1,372

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

(1) Robust standard errors are clustered at the school level and shown in parenthesis.
(2) The baseline and end-line data are pooled in columns (I) and (II). The panel data used for columns (III) and (IV) is constructed from the baseline and end-line data.
(3) All test scores at each survey round are normalized by the mean and the standard deviations of the baseline scores of 4th grade students. The maximum scale (full mark) of math test score is 56.

61 Cost-effective analysis is conducted for the sampled schools, following the methodology presented by J-PAL (Bhula, Mahoney and Murphy, 2020; Dhaliwal et al., 2014). Since the data on student participation rates in remedial activities at the regional level is not available, this study limits the cost-effectiveness analysis to the sampled schools that had 4th grade students. There were thus 28 schools in total. As only the overall cost data of scaling up PMAQ in the Tillaberi region is available, this study calculated unit training costs and the transportation costs of workbooks per school. The unit cost of workbooks was calculated per student, supposing that each student utilized three workbooks in total during the remedial activities.

The cost-effectiveness is measured as the ratio of the aggregated impact of the project (the average impact on student learning per student multiplied by the number of students impacted) to the aggregated cost of implementing the project. The cost-effectiveness is presented as the total standard deviations gained across the sample per 100 USD spent.

The cost-effectiveness of the additional interventions of PMAQ, the total standard deviations gained across the sample per 100 USD spent, is estimated to be 2.14. The level of cost-effectiveness of the additional interventions in PMAQ is comparable to programs cited in Kremer et al. (2013).

Spider charts in Appendix 16 (Graph D-1 to D-3) display the before/after comparison of the correct response rates for each test item among 4th grade students in the panel data. The correct response rates for each item improved from baseline to end-line, but the rates are still low in items numbers 31 and 32 (two-digit subtraction with borrowing), and items from numbers 52 through 56 (three to four-digit addition with carrying, and subtraction with borrowing). The test items that exhibit an improved correct response rate differ according to the baseline score level. For example, students with a baseline score of less than the median correctly answered more single addition and subtraction items in the end-line survey (Items No. 1 through 12 on Graph D-2). Students with a baseline score over the median correctly answered more test items involving single-digit multiplication (Items No. 36 through 44) and three-digit addition and subtraction without borrowing (Items No. 45 through 51 on Graph D-3).

4.4.2 Testing the parallel trends assumption

The trends in average scores of the 4th and 5th grade student groups from the pre-baseline to the baseline surveys are nearly parallel (Graph 1). The parallel trends are checked by estimating equation (3) using the panel data of pre-baseline and baseline scores. The parameter estimate for λ' is -0.062 standard deviations, which is not statistically significant (Table 16 (I)).[62] This shows that the parallel trends assumption holds for identifying the impact of PMAQ through the DD of the test scores of 4th and 5th grade students.

Next, the trends in math learning for 5th and 6th grade students who were not targeted by remedial activities using math workbooks are checked. If the trends in the test scores of the 5th and 6th grade students are parallel, the results support the equal trends assumption between 4th and 5th grade students in the absence of PMAQ. The results from equation (4) and (5) are shown in Table 16 (II) through (IV). The parameter estimates for λ'' are 0.068 standard deviations using the pooled data and 0.017 standard deviations using the panel data. Neither of them is statistically significant. Those results support the idea that the DD of the 4th and 5th grade student test scores accurately identifies the impact of PMAQ on math learning.

---

[62] Most 4th grade students were studying in the 3rd grade when the pre-baseline survey was conducted. As explained in section 4.(2), the maximum scale for 3rd grade students is 44.

Table 16. Regression results of parallel trend test

	(I) **4th / 5th** Panel: 2t Model (3)	(II) **5th / 6th** Pooled Model (4)	(III) **5th / 6th** Panel: 2t Model (4)	(IV) **5th / 6th** Panel: 2t Model (5)
Grade 4 (Treatment) × Pre-Treatment (Baseline) (λ’)	-0.062			
	(0.137)			
Grade 5 × Post (λ’’)		0.068	0.017	0.017
		(0.091)	(0.089)	(0.125)
Grade 5		-0.342**	-0.262	
		(0.136)	(0.166)	
Pre-Treatment (Baseline)	0.346***			
	(0.091)			
Post		0.389***	0.415***	0.415***
		(0.075)	(0.075)	(0.105)
Female		-0.156*	-0.151	
		(0.084)	(0.104)	
School fixed effect (dummy)	No	Yes	Yes	No
Student fixed effect (dummy)	Yes	No	No	Yes
Adjusted R-Squared	0.808	0.321	0.307	0.813
Observations	906	2,041	1,426	1,426

***p < 0.01, **p < 0.05, *p < 0.1

(1) Robust standard errors are clustered at the school level and shown in parenthesis.
(2) The panel data used for column (I) is composed from the pre-baseline and baseline data. The panel data used for column (II) to (IV) are composed from the baseline and end-line data.
(3) The maximum scale (full mark) of math test scores in column (I) is 44. The test scores are normalized by the mean and standard deviations of the 4th grade student group in the pre-baseline survey.
(4) The maximum scale (full marks) of math test score from columns (II) to (IV) is 56. The test scores are normalized by the mean and standard deviations of 5th grade students in the baseline survey.

4.4.3 Robustness check of the estimated impact

To check the robustness of the estimated impact of PMAQ further, the interaction terms of X_i and $Post_t$, and those of X_j and $Post_t$ are added in equations (1) and (1-a). The estimates of λ from theses equations remain almost at the same level as reported in the previous sub-section (Table E-1 (I) to (III) in Appendix 17).

Given the structure of the math test, students with higher baseline scores find it harder to gain more scores at the end-line survey than students with lower baseline scores. Since the average baseline scores of 4th grade students are below those of 5th grade students, the differences in the baseline score level might engender a bias in estimating the impact of PMAQ. To check such potential bias, this study conducted a regression with the sub-samples in the same range of baseline test scores from 10 to 37 score points, equivalent to the 2nd and 3rd quartile of the 4th grade baseline scores. The sub-sample of 5th grade students in the bandwidth represents around half of the panel data sample (Graph D-4 and D-5 in Appendix 16).[63] The test scores are standardized with the mean and the standard deviations of baseline test scores of the whole sample of 4th grade students. The estimates for λ using the sub-sample are in the range of 0.27 to 0.28, which is slightly smaller than the values from the whole sample but still statistically significant (Table E-2 in Appendix 17).

Although the distribution of math workbooks was limited to 1st through 4th grade students, some schools also organized supplementary classes for 5th and 6th grade students. The supplementary classes for 5th and 6th grade students may have been aimed at numbers and the four basic math operations without using the math workbooks. To check the robustness of the estimated impact of PMAQ, this study conducted a regression with 4th grade students and the sub-samples of 5th grade students from those schools that organized supplementary classes for this grade. The test scores are standardized with the mean and standard deviations of the baseline test scores of the whole sample of 4th grade students. The estimate for λ using the limited sample is 0.34, which is slightly below the values from the whole sample but still statistically significant (Table E-3 in Appendix 17). If the 5th grade students participated in supplementary classes of the same duration as the 4th grade students, and with a focus on numbers and the four basic operations but without the use of math workbooks, the estimate would represent the sole impact of the distribution of math workbooks.

4.4.4 Heterogeneous impacts

This study analyzes the heterogeneity of the impacts of PMAQ on student math learning with respect to the baseline score level and gender. If the impact is larger on students with lower baseline test scores, this indicates that PMAQ helps those students to catch up their math learning. If the impact is larger on girls than on boys, it shows that PMAQ reduces gender disparities in math learning. As shown in Graph D-6 in Appendix 16, the average baseline score for girls is lower than for boys.

---

[63] The percentage of 5th grade students whose baseline score was 9 score points or below is 9 percent of the original two-period panel data sample. The percentage of 5th grade students with baselines below 37 score points is 57 percent of the two-period panel data sample.

Equation (1-a) is expanded for the heterogeneity analysis as:

$$Y_{ijt} = \alpha_j + \beta_1\, Grade4_i + \beta_2\, Post_t + (\lambda_0 + \lambda_1\, Female_i)\,(Grade4_i \times Post_t) + X_i\gamma_1 + \varepsilon_{ijt} \quad \text{(6-1), and}$$

$$Y_{ijt} = \alpha_j + \beta_1\, Grade4_i + \beta_2\, Post_t + (\lambda_0 + \lambda_{1L}\, Lowest\_Score_i + \lambda_{1M}\, Median\_Score_i)\,(Grade4_i \times Post_t) + X_i\gamma_1 + \varepsilon_{ijt} \quad \text{(6-2).}$$

The heterogeneous impact is estimated with the balanced panel data. In equation (6-1), parameter λ_1 identifies the treatment effect additional to girls in comparison to boys. In equation (6-2), the sample of students is categorized into the following three groups: the lowest baseline score group (the 1st quartile of the score of each grade); the median baseline score group (the 2nd and 3rd quartiles of the score of each grade); and the highest baseline score group (the 4th quartile of the score of each grade). This study then defines $Lowest\_Score_i$ as the dummy variable that takes the value 1 if the test score of student *i* falls into the lowest baseline score group, and $Median\_Score_i$ as the dummy variable which takes the value 1 if the baseline test score of student *i* falls into the median baseline score group. Parameters λ_{1L} and λ_{1M} will identify the additional treatment effect for students in different groups in comparison to the highest baseline score group.

Although the female dummy and lower baseline score dummy are positively correlated, it is possible to include both in the same equation. The heterogeneity impact model in equation (6) can be expanded as:

$$Y_{ijt} = \alpha_j + \beta_1\, Grade4_i + \beta_2\, Post_t + (\lambda_0 + \lambda_{1L}\, Lowest\_Score_i + \lambda_{1M}\, Median\_Score_i + \lambda_2\, Female_i)(Grade4_i \times Post_t) + X_i\gamma_1 + \varepsilon_{ijt}\,, \quad \text{(7), and}$$

$$Y_{ijt} = \alpha_j + \beta_1\, Grade4_i + \beta_2\, Post_t + (\lambda_0 + \lambda_{1L}\, Lowest\_Score_i + \lambda_{1M}\, Median\_Score_i + \lambda_2\, Female_i + \lambda_{3L}\, Lowest\_Score_i \times Female_i + \lambda_{3M}\, Median\_Score_i \times Female_i)(Grade4_i \times Post_t) + X_i\gamma_1 + \varepsilon_{ijt}\,, \quad \text{(8),}$$

from which it is possible to test whether there is an interaction effect between the two types of disadvantages combined through examining parameter λ_3.

The impact of PMAQ is larger for the lower baseline score groups (Table 17-1 (II), (IV), and (V)). The estimate for λ_{1L} is 0.55 standard deviations, and that for λ_{1M} is 0.20 standard deviations, even after controlling for the heterogeneous impact by gender (Table 17-1 (V)). The result indicates that PMAQ helped children with lower baseline scores to catch up with the higher baseline score group. In contrast, the additional impact for girls is not statistically significant (Table 17-1 (IV) and (V)).

To examine the robustness of the heterogeneous impacts, the interaction terms of X_i and $Post_t$ are added to equations (6) to (8). For instance, when an interaction term of the female dummy and $Post_t$ is added, λ_2 in the equations represents the additional impact of PMAQ on the 4th grade girls in comparison to the 5th grade girls, independent of the additional impact on boys. As noted in the previous section, since the average baseline scores of 4th grade students are below those of 5th grade students, the difference in the baseline score level might engender a bias in estimating the impact of PMAQ. For this reason, $Lowest\_Score_i$ is replaced with $Lowest\_Score'_i$ (taking the value of 1 when the test score of student i is 0 to 9 score points, that is the range of the 1st quartile baseline score of 4th grade students), and also $Median\_Score_i$ with $Median\_Score'_i$ (taking the value of 1 when the test score of student i is 10 to 37 score points, that is the range of the 2nd and 3rd quartiles of 4th grade students) in equations (6) to (8).

The estimation results are reported in Table 17-2. The impact on the highest score group (λ_0) is positive and statistically significant at the 1 percent level (Table 17-2 (II)). Fourth grade students in the highest baseline test scores group learned math better than the 5th grade students on the same bandwidth of the test score. On the other hand, the impact on the lowest test scores group (λ_{1L}) is positive but not statistically significant (Table 17-2 (II) and (V)). These results show that the lowest test score group of 4th grade students as well as that of 5th grade students caught up with the highest score group of each grade. In terms of the gender disparity in math learning, the average baseline score of girls is lower than boys, and the time trend in learning among 5th grade girls is negative (Table 17-2 (III), (IV) and (V)), which indicates that the gender disparity in learning outcomes tends to increase over time. The estimated value λ_2 of the heterogeneous impact by gender is positive and statistically significant at the 5 percent level (Table 17-2 (III) to (V)), which demonstrates that girls in the 4th grade could learn math better than girls in the 5th grade.

Table 17-1. Heterogeneous impacts of PMAQ on math learning

	(I) **4th/5th** Panel: 2t Model (1-a)	**(II)** **4th/5th** Panel: 2t Model (6)	**(III)** **4th/5th** Panel: 2t Model (6)	**(IV)** **4th/5th** Panel: 2t Model (7)	**(V)** **4th/5th** Panel: 2t Model (8)
Grade4 (Treatment) ×Post: DD (λ_0)	0.387*** (0.076)	0.142 (0.096)	0.347*** (0.085)	0.118 (0.109)	0.080 (0.132)
DD×Lowest score (λ_{1L})		0.555*** (0.128)		0.550*** (0.128)	0.490*** (0.171)
DD×Median score (λ_{1M})		0.207*** (0.071)		0.207*** (0.070)	0.300*** (0.102)
DD×Female (λ_2)			0.074 (0.053)	0.047 (0.058)	0.123 (0.093)
DD×Lowest score ×Female (λ_{3L})					0.074 (0.170)
DD×Median score ×Female (λ_{3M})					-0.180 (0.121)
Grade 4 (Treatment)	-0.511*** (0.109)	-0.578*** (0.051)	-0.584*** (0.052)	-0.578*** (0.051)	-0.578*** (0.051)
Post	0.394*** (0.055)	0.394*** (0.055)	0.394*** (0.055)	0.394*** (0.055)	0.394*** (0.055)
Lowest score		-2.108*** (0.066)	-1.972*** (0.050)	-2.107*** (0.066)	-2.176*** (0.084)
Median score		-0.891*** (0.044)	-0.842*** (0.031)	-0.890*** (0.044)	-0.900*** (0.062)
Female	-0.160** (0.074)	-0.084* (0.050)	-0.100* (0.051)	-0.096* (0.049)	-0.150*** (0.053)
Lowest score× Female					0.150 (0.107)
Median score× Female					0.027 (0.075)
School fixed effect	Yes	Yes	Yes	Yes	Yes
Student fixed effect	No	No	No	No	No
Adj. R^2	0.515	0.794	0.788	0.793	0.794
Observations	1,372	1,372	1,372	1,372	1,372

***p < 0.01, **p < 0.05, *p < 0.1

(1) Robust standard errors are clustered at the school level and shown in parenthesis.

(2) All test scores are normalized with the mean and the standard deviations of the baseline scores of 4th grade students. The maximum scale (full mark) of math test score is 56.

(3) The lowest score dummy takes value 1 for students who obtained baseline scores below the 1st quartile score of the 4th grade student baseline scores. Median score dummy takes value 1 for students who obtained baseline scores higher than the 1st quartile score of 4th grade student baseline scores, and below the 3rd quartile of the scores.

Table 17-2. Robustness check of the heterogeneous impacts of PMAQ

	(I) **4th/5th** Panel: 2t Model (1-c)	**(II)** **4th/5th** Panel: 2t Model (6-a)	**(III)** **4th/5th** Panel: 2t Model (6-a)	**(IV)** **4th/5th** Panel: 2t Model (7-a)	**(V)** **4th/5th** Panel: 2t Model (8-a)
Grade4 (Treatment) ×Post: DD (λ_0)	0.390^{***} (0.077)	0.210^{***} (0.068)	0.165^{**} (0.067)	0.116 (0.075)	0.091 (0.091)
DD×Lowest score (λ_{1L})		0.305 (0.182)		0.302 (0.184)	0.288 (0.188)
DD×Median score (λ_{1M})		0.015 (0.077)		0.016 (0.080)	0.089 (0.118)
DD×Female (λ_2)			0.195^{**} (0.091)	0.193^{**} (0.091)	0.240^{**} (0.099)
DD×Lowest score ×Female (λ_{3L})					-0.051 (0.394)
DD×Median score ×Female(λ_{3M})					-0.131 (0.189)
Grade 4 (Treatment)	-0.513^{***} (0.109)	-0.108^{**} (0.042)	-0.113^{**} (0.043)	-0.109^{**} (0.042)	-0.110^{**} (0.042)
Post	0.411^{***} (0.056)	0.142^{***} (0.042)	0.167^{***} (0.036)	0.188^{***} (0.045)	0.174^{***} (0.033)
Lowest score		-2.424^{***} (0.060)	-2.427^{***} (0.061)	-2.425^{***} (0.060)	-2.471^{***} (0.052)
Median score		-1.227^{***} (0.028)	-1.228^{***} (0.028)	-1.228^{***} (0.028)	-1.242^{***} (0.048)
Female	-0.141^{*} (0.075)	-0.100^{**} (0.038)	-0.100^{**} (0.038)	-0.101^{**} (0.038)	-0.135^{***} (0.028)
Lowest score ×Female					0.101 (0.067)
Median score ×Female					0.035 (0.067)
Post×Lowest score		0.517^{***} (0.141)	0.719^{***} (0.102)	0.511^{***} (0.145)	0.468^{***} (0.128)
Post×Median score		0.482^{***} (0.060)	0.481^{***} (0.042)	0.481^{***} (0.058)	0.514^{***} (0.071)
Post×Female	-0.037 (0.040)	-0.052 (0.045)	-0.148^{**} (0.073)	-0.153^{**} (0.070)	-0.121 (0.077)
Post×Lowest score ×Female					0.125 (0.383)
Post×Median score ×Female					-0.075 (0.122)
School fixed effect	Yes	Yes	Yes	Yes	Yes
Student fixed effect	No	No	No	No	No
Adj. R^2	0.514	0.788	0.788	0.789	0.789
Observations	1,372	1,372	1,372	1,372	1,372

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

(1) Robust standard errors are clustered at the school level and shown in parenthesis.

(2) All test scores are normalized with the mean and the standard deviations of baseline scores of 4th grade students. The maximum scale (full marks) of math test score is 56.

(3) Lowest score dummy takes value 1 for students who obtained baseline scores below the 1st quartile score of 4th grade student baseline scores. Median score dummy takes value 1 for students who obtained baseline scores more than the 1st quartile scores of 4th grade student baseline scores, and below the 3rd quartile of the scores.

(4) Models (6-a) to (8-a) are expanded from models (6) to (8) by adding the interaction term of the end-line dummy and student characteristics.

4.5 Discussion

Through a package of different interventions, PMAQ improved the math learning of primary 4th grade students in the Tillaberi region. The impact was larger for students with lower baseline scores. PMAQ employed a self-learning style that was supported by activity facilitators. The self-learning style of extra-curricular remedial activities in PMAQ is similar to computer-aided learning programs. In a computer-aided learning program, students follow instructions on the screen and learn math individually with the help of facilitators. The computer-aided learning program can deliver individually customized content that matches student learning levels. Muralidharan et al. (2019) demonstrate that a computer-aided learning program improves learning outcomes in language and math for students at the lower secondary level in urban areas in India. Though PMAQ employs paper-based learning using math workbooks, the workbooks were designed to respond to the various proficiency levels of students (Hara et al., 2020). While the two types of learning materials are different, they are similar in that students individually learn math by accessing learning materials with the support of facilitators. In sub-Saharan Africa, where access to electricity is still limited, paper-based self-learning material remains an important tool.

Beasley and Huillery (2016) investigated the impact of the provision of school grants for SMCs in Niger. While parental involvement in school management was enhanced by the provision of a school grant, the intervention did not improve student learning. The study suggests that it is important to engage teachers when seeking to improve student learning. Banerjee et al. (2010) studied the strategy of community engagement in student learning through an experiment in India and showed that in order to improve student learning, it is necessary to provide a specific course of action for the local community. PMAQ targeted stakeholders both in schools and the local communities to improve student learning through extra-curricular remedial activities. SMC organized a community general assembly where it shared math assessment test scores with parents, teachers, and community members; then the committee facilitated the discussion on the development of joint action plans (Hara et al., 2020).

While PMAQ helped primary 4th grade students improve math learning, this study found that these students still struggled with solving the items of two-digit subtraction with borrowing, three to four-digit addition with carrying, and subtraction with borrowing. An effective way of overcoming this difficulty would be to strengthen the pedagogical intervention in PMAQ by introducing a different pedagogical method. The EPT project is experimenting with integrating the method of "Teaching at the Right Level (TaRL)," which has been developed by the Indian NGO "Pratham," into PMAQ. Under the TaRL method, students are provided with different types of math activities through which they can advance and strengthen their conceptual understanding of the four arithmetic operations.

PMAQ helped students with lower baseline scores improve math learning; however, it was not sufficient to decrease the gender disparity in math learning. This study found that 4th grade girls who learned by using the workbook were able to improve their math learning better than 5th grade girls who did not; however, the gender disparity in basic math learning still remains. In Niger, the percentage of girls who reach the last grade of primary education is 65 percent, which is lower than that of boys by 13 percentage points in 2016 (UIS, 2020). Poor learning achievement is a major reason for girls dropping out of primary education (Perlman et al., 2016); thus PMAQ has room to reduce the gender disparity in learning.

4.6 Conclusions

The scaling-up of PMAQ in Niger's Tillaberi region was aimed at helping students improve their math learning through extra-curricular remedial activities over a three-month period. PMAQ includes two components: a one-day training course for SMC presidents and secretaries to strengthen the capacity of SMCs to organize remedial activities in math, and the distribution of a set of math workbooks to students, accompanied by a two-day training for activity facilitators. Due to budget constraints, the distribution of math workbooks was limited to primary 1st to 4th grade students. Focusing on the difference in the intervention between 4th and 5th grade students, this study investigated the impact of PMAQ on student math learning using three rounds of survey data. The average impact on math learning outcomes is estimated to be 0.36 to 0.38 standard deviations. The impact is larger for students with lower baseline scores, which indicates that the remedial activity helped students left behind in math to catch up. Considering that the period of extra-curricular remedial activities was only three months, the progress in math learning is encouraging.

After scaling up PMAQ in Niger in 2018, the JICA-supported project upgraded PMAQ to cover basic reading skills by employing an effective pedagogical approach called "Teaching at the Right Level (TaRL)" in Madagascar (Maruyama et al., 2021). In the remedial activities of the upgraded model, students learn basic reading and then study how to solve math problems posed in a text. By integrating the basic reading component, PMAQ support students in developing their skills in applying the knowledge of mathematics in their daily lives.

In developing countries, the mismatch between curricula and student learning levels expands as students progress through the grades. Although most students are left behind, teachers are required to finish the prescribed syllabus. As a result, many children in sub-Saharan Africa finish or drop out of primary education without mastering the foundation skills including math. The scaling up of PMAQ in Niger suggests that once children have a chance to learn in a manner that matches their learning levels, they can advance their math learning. To overcome the learning crisis in sub-Saharan Africa, the governments in the region should take policy measures that help children master the foundation skills, and the development partners should support the initiative.

Appendix 13: Item analysis of the baseline math test

Appendix 14: Panel data construction and gender identification

Appendix 15: Attrition analysis of student data

Appendix 16: Graphs and table of math test results

Appendix 17: Tables on the robustness check of the estimated impact of PMAQ

References

Abadzi, Helen, and Stavri Llambiri. 2011. "Selective Teacher Attention in Lower-Income Countries: A Phenomenon Linked to Dropout and Illiteracy?" *Prospects*, *41*: 491–506.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics*, *122*(3): 1235–1264.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster and Stuti Khemani. 2010. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India." *American Economic Journal: Economic Policy*, *2*(1), 1-30.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobini Mukerji, Marc Shotland. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives*, *31*(4): 73-102.

Banerji, Rukmini. 2000. "Poverty and Primary Schooling: Field Studies from Mumbai and Delhi." *Economic and Political Weekly*, *35*(10): 795-802.

Bashir, Sajitha, Marlaine Lockheed, Elizabeth Ninan, and Jee-Peng Tan. 2018. *Facing Forward: Schooling for Learning in Africa*. Washington DC: World Bank.

Beasley, Elizabeth, and Elise Huillery. 2016. "Willing but Unable? Short-term Experimental Evidence on Parent Empowerment and School Quality." *The World Bank Economic Review*, *31*(2): 531–552.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane. 2017. "Enrollment without Learning: Teacher Effort, Knowledge and Skill in Primary Schools in Africa." *Journal of Economic Perspectives*, *31*(4): 185-204.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabub, Alice Ng'ang'a, Justin Sandefur. 2018. "Experimental evidence on scaling up education reforms in Kenya." *Journal of Public Economics*, *168*: 1–20.

Bhula, Radhika, Meghan Mahoney, and Kyle Murphy. 2021. Conducting cost-effectiveness analysis (CEA). Retrieved from https://www.povertyactionlab.org/resource/conducting-cost-effectiveness-analysis-cea

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. (2014). Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education. In Paul Glewwe (Ed.). *Education Policy in Developing Countries* (pp. 285-338). Chicago: University of Chicago Press.

Evans, David K., and Anna 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.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children left behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics, 1*(1): 112-35.

Hara, Masahiro, Takao Maruyama, Akiko Kageyama, and Nobuhiro Kunieda. 2020. "Quality Learning through Community-wide Collaboration: A Methodology to Overcome the "Learning Crisis" in Niger." In M. Nishimura (Ed.). *Community participation with schools in developing countries: Towards equitable and inclusive basic education for all*. Routledge Research in Educational Equality and Diversity Series. New York & Oxford: Routledge.

Honda, Shunichiro, and Hiroshi Kato. 2013. "Scaling up in education: School-based management in Niger." In Laurence Chandy, Akio Hosono, Homi Kharas, and Johannes F. Linn. (Eds.). *Getting to scale: How to bring development solutions to millions of poor people.* Washington, DC: Brookings Institution.

JICA 2017. *Final Evaluation Report on School for All: The project on support to educational development through community participation*. Tokyo: Japan International Cooperation Agency.

Kerwin, Jason, T., and Thornton, Rebecca L. 2020. "Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures." *Review of Economics and Statistics*: 103 (2): 1-45.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science*, *340,* 297.

Kunieda, Nobuhiro, Takao Maruyama, Akiko Kageyama, and Masahiro Hara. 2020. "Educational Development through Community-wide Collaboration: How to Establish a Sustainable Community-wide Initiative to Improve Education." In M. Nishimura (Ed.), *op.cit.*

Maruyama, Takao, Kengo Igei and Seiichi Kurokawa. 2021. "Community-wide support for primary students to improve learning: Empirical evidence from Madagascar." RISE Annual Conference 2021.

Ministry of Primary Education in Niger. 2016. *Statistiques De L'éducation De Base Et Alphabétisation Annuaire: Annuaire 2015-2016*. Niamey: The Ministry of Primary Education in Niger.

Ministry of Primary Education in Niger. 2018. *Data from BILAN* [Situation de la mise en œuvre du PMAQ dans les écoles]. Niamey: The Ministry of Primary Education in Niger.

Ministry of Primary Education in Niger. 2019. *Statistiques De L'éducation De Base Et Alphabétisation Annuaire*. Niamey: The Ministry of Primary Education in Niger.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting education? Experimental evidence on technology-aided instruction in India." *American Economic Review*, *109*(4) : 1426-60.

PASEC. (2016). *PASEC 2014 – Performances du système éducatif nigérien: Compétences et facteurs de réussite au primaire*. Dakar: PASEC. CONFEMEN.

Perlman, Daniel, Fatima Adamu, Quentin Wodon. 2018. "Why Do Adolescent Girls Drop Out of School in Niger? A Combined Quantitative and Qualitative Analysis." *Marché et organisations*, *2* (32): 179-194.

Pratham 2020. Teaching at the Right Level: From concern with exclusion to challenges of implementation. Background paper prepared for the 2020 Global Education Monitoring Report.

Pritchett, Lant, and Amanda Beatty. 2015. "Slow Down, You Are Going Too Fast: Matching Curricula to Student Skill level." *International Journal of Educational Development*, *40*: 276-288.

Robinson, David. 2019. *fuzzyjoin: Join Tables Together on Inexact Matching. R package version 0.1.5*. Retrieved from https://CRAN.R-project.org/package=fuzzyjoin.

Sabarwal, Shwetlena, David K. Evans, Anastasia Marshak. 2014. "The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone." *Policy Research Working Paper*, 7021. Washington DC: The World Bank.

The "Ecole pour tous" project. (2018).Résolutions à l'atelier de lancement de la vulgarisation du PMAQ Math dans la région de Tillabéri tenu le 8 mars 2018 à Kollo. Tillaberi: The "Ecole pour tous" project.

World Bank. 2017. *Republic of Niger: Niger service delivery indicators education 2015*. Washington DC: World Bank.

Item analysis of the baseline math tests

The items in the math tests are analyzed using the baseline sample of 4th grade students (N= 545) from the pooled data. The results are shown in Table A.1. "Correct response rate" is the percentage of 4th grade students who answered each item correctly. "Item-total correlation" is the correlation between a vector of student responses (correct answer=1; wrong answer=0) to the item *n* and a vector of the total scores excluding the item *n*. Then, the item difficulty and the item discrimination are estimated, using the two-parameter logistic model of the item response theory.[i] Cronbach's Alpha is also calculated for each section of the math test. The values are 0.94 (section (i)), 0.95 (section (ii)), 0.90 (section (iii)), 0.93 (section (iv)), and 0.88.

These results indicate that the items of the math test used in the baseline survey were properly able to capture the different levels of student math achievement in the basic four arithmetic operations.

Table A.1. The result of the item analysis of the math test

	Item	Correct response rate (percentage)	Item-Total Correlation	Parameter of item difficulty	Parameter of item discrimination
Section (i) single-digit addition					
1	3+1	85.3	0.45	-1.54	2.17
2	1+4	82.8	0.51	-1.32	2.75
3	3+0	76.3	0.50	-1.14	1.74
4	4+4	68.1	0.62	-0.70	2.22
5	5+4	71.6	0.65	-0.79	3.35
6	7+2	71.7	0.64	-0.82	2.75
7	8+2	73.2	0.60	-0.90	2.42
8	3+7	64.4	0.63	-0.56	2.19
9	5+5	69.4	0.65	-0.73	2.66

[i] This study used the R package IRTOYS to estimate the parameters of the two-parameter logistic model.

10	8+7	60.4	0.66	-0.40	2.22
11	5+9	56.9	0.66	-0.27	2.04
12	6+6	55.8	0.71	-0.22	2.53
Section (ii) single-digit subtraction					
13	3-1	58.5	0.66	-0.34	1.96
14	5-2	57.2	0.71	-0.27	2.47
15	4-4	48.4	0.71	0.04	2.31
16	2-0	52.8	0.70	-0.12	2.20
17	9-6	46.2	0.75	0.12	2.74
18	8-8	46.2	0.74	0.12	2.73
19	10-9	44.2	0.74	0.20	2.61
20	10-3	51.4	0.69	-0.06	2.05
21	10-6	46.8	0.58	0.12	1.44
22	19-4	32.7	0.69	0.63	2.34
23	15-6	36.9	0.72	0.47	2.57
24	14-8	33.0	0.68	0.62	2.34
Section (iii) two-digit addition including problems with carrying, and subtraction including problems with borrowing					
25	33+24	61.3	0.63	-0.46	1.89
26	52+7	43.3	0.64	0.25	1.75
27	8+55	28.6	0.62	0.82	2.01
28	79+16	26.1	0.63	0.92	2.16
29	94-42	38.0	0.69	0.44	2.15
30	88-58	35.6	0.65	0.54	1.97
31	74-47	8.8	0.43	1.85	2.22
32	80-34	6.6	0.34	2.21	1.82
33	69+75	20.6	0.62	1.11	2.65
34	58+63	23.5	0.64	0.99	2.57
35	92+8	23.9	0.63	0.99	2.42
Section (iv) single-digit multiplication					
36	2×4	45.9	0.66	0.15	1.78
37	3×1	41.7	0.63	0.32	1.61
38	5×7	30.5	0.66	0.74	2.08
39	8×3	31.6	0.64	0.71	1.89
40	9×2	25.1	0.55	1.07	1.55
41	7×5	31.6	0.64	0.71	1.90
42	0×7	26.1	0.40	1.34	0.95

43	6×6	26.8	0.61	0.94	1.79
44	9×7	21.5	0.51	1.29	1.47
Section (v) three-to-four-digit addition including problems with carrying, and subtraction including problems with borrowing					
45	627+370	45.7	0.66	0.15	1.86
46	173+807	24.2	0.65	0.96	2.55
47	744+187	21.1	0.63	1.09	2.65
48	488+557	16.7	0.57	1.28	2.74
49	303-201	28.4	0.67	0.79	2.52
50	617-214	29.7	0.66	0.75	2.32
51	788-780	26.4	0.63	0.90	2.18
52	630-517	5.7	0.37	1.98	2.87
53	637-539	1.3	0.19	2.76	2.81
54	402-353	3.1	0.31	2.06	4.69
55	1976+9587	10.3	0.46	1.68	2.50
56	9101-6609	2.2	0.26	2.31	3.72

Appendix 14

Panel data construction and gender identification

A local NGO conducted three rounds of survey in March 2017 (pre-baseline), December 2017 (baseline), and June 2018 (end-line). Since the three rounds of survey in this study were not designed to track the same students, this study constructed panel data for 4th through 6th grade students ex post by manually linking data in the three rounds of the survey. As student names were typed for each round of the survey, this had the potential to cause inconsistencies of spelling across the three rounds. To overcome any inconsistencies in the spelling of student names, the fuzzy joining method of Robinson (2019) is utilized.

Specifically, this study took four steps to construct panel data from the three rounds of survey data. First, accent marks above letters in student names were removed, and student and school names were converted into lowercase letters.

Next, this study assigned identification numbers to each student in the end-line data. Personal names in Niger generally consist of an individual's first name, father's first name, and grandfather's first name. Some of the student names in the registry had all of these, but others had missing values for the grandfather's name. Thus, this study fuzzy joined the end-line data with the baseline data grade by grade, using the letters of the student's first name, his/her father's name, and the name of the school. This study allowed a degree of fuzziness up to one-letter difference in the child's first name and one-letter difference in the father's name for each student. If there was more than one student in the baseline data who was fuzzy joined with a student in the end-line data, exact matching was tried to exclude the duplication, using first name and father's name. When there were students with exactly the same name attending the same grade in the same school, this study used grandfather's name and matched it again. In case there were still some students with the same name, this study excluded them from the analysis.[ii]

[ii] Five pairs of students had the same first and family name and attended the same school. This study excluded them from the panel data sample.

As a result, the panel dataset of two data points (baseline and end-line) includes 1,070 students from 4th through 6th grade.

Third, this study fuzzy joined the panel data (baseline and end-line) with pre-baseline data for the corresponding previous grade or the same grade during the 2016-17 school year. For example, this study fuzzy joined 4th grade students from the 2017-18 school year with 3rd grade students from the previous school year.[iii] Since students might have repeated the grade, this study fuzzy joined the remaining unmatched data with 4th grade students from the 2016-17 school year. The panel dataset of three data points contains 717 students.

The gender of the students was not recorded in any of the three survey rounds. This study recovered the student's gender information based on the typical correspondence between first name and gender. There were approximately 700 different first names in the pooled data from the baseline and end-line surveys. The EPT project provided the association between first names and gender without difficulty as boys' and girls' names are rarely gender-ambiguous in Niger. In the process of panel data construction, this study recovered the gender information in the end-line data, then fuzzy matched this with the baseline data. As noted above, in the process of fuzzy matching, this study allowed a one-letter difference in first name and a one-letter difference in father's name for each student.

In cases where there was a one-letter difference in the student's first name, this could have caused an inconsistency in the gender identification. There were six students in total who had a first name with a one-letter difference in the matched sample, which caused inconsistency in gender identification, i.e., *abassa* (girl) and *abass* (boy), *hassana* (girl) and *hassane* (boy), *kanwara* (girl) and *kanwari* (boy), *nomie* (girl) and *nomi* (boy), *rachida* (girl) and *rachid* (boy), *imrana* (girl) and *imran* (boy). This study excluded these students from the panel data analysis.

[iii] Since this study fuzzy joined different waves of survey data, this study came across cases in which the name of a student was different between the end-line and baseline data by one letter. In such cases, this study used the name in the end-line to link them with the pre-baseline data.

Appendix 15

Attrition analysis of student data

As explained in Appendix 2, the panel data was constructed through fuzzy-joining the end-line data with the baseline data. Of the 545 4th grade students in the baseline survey, 357 remained in the panel data (attrition of 188 students in the panel data). Of the 524 5th grade students in the baseline survey, 329 remained in the panel data (attrition of 195 students in the panel data). To check differential attrition occurred between the 4th and 5th grades, the attrition dummy is regressed on the 4th grade dummy, female dummy, student baseline scores, and school characteristics. The regression results are presented in Table C.1. While slightly higher attrition is observed for students with lower baseline scores, differential attrition did not occur between the 4th and 5th grade students in the panel data constructed by the baseline and end-line data (Column (1)). Similarly, this study checked whether differential attrition occurred between the 5th and 6th grade. Differential attrition did not occur between the 5th and 6th grade students (Column (2) in Table C.1).

Table C.1. Student attrition analysis on the panel data constructed from

the baseline and end-line data

	Grade 4 & 5		Grade 5 & 6	
	(1)-1	**(1)-2**	**(2)-1**	**(2)-2**
(Intercept)	0.408***		0.342***	
	(0.050)		(0.042)	
Grade 4	-0.069	-0.026		
	(0.048)	(0.053)		
Grade 5			0.033	0.010
			(0.041)	(0.038)
Female	-0.024	-0.014	-0.021	-0.021
	(0.035)	(0.030)	(0.034)	(0.029)
Z score baseline	-0.058***	-0.051***	-0.028	-0.040**
	(0.017)	(0.016)	(0.020)	(0.016)
Franco-Arab school	0.050		-0.071	
	(0.060)		(0.101)	
Basic reading pilot school	0.019		0.395***	
	(0.153)		(0.038)	
School fixed effect	No	Yes	No	Yes
Adj. R^2	0.012	0.415	0.017	0.423
Num. obs.	1,066	1,066	1,092	1,092

***p < 0.01; **p < 0.05; *p < 0.1

(1) Robust standard errors are clustered at the school level and shown in parenthesis.
(2) In Column (1), all test scores for each survey round are normalized by the mean and standard deviations of the baseline scores of the 4th grade students. The maximum scale (full mark) of math test score is 56.
(3) In Column (2), all test scores for each survey round are normalized by the mean and standard deviations of the baseline scores of the 5th grade students. The maximum scale (full mark) of math test score is 56.

In section 4.4.2 of Chapter 4, this study checked the parallel trends before the interventions. As explained in Appendix 2, this study fuzzy-joined the panel data (constructed by the baseline and end-line data) with the pre-baseline survey. Of the 357 students in the 4th grade group of the panel data (constructed by the baseline and end-line data), 236 students in the pre-baseline data were fuzzy-joined (attrition of 121 students). Of the 329 students in the 5th grade group in the panel data (constructed by the baseline and end-line data), 217 students in the pre-baseline data were fuzzy joined (attrition of 112 students). To check differential attrition occurred between the 4th grade and 5th grade groups in the process of fuzzy joining the panel data with the pre-baseline data, the attrition dummy is regressed on the 4th grade dummy, female dummy, student baseline scores, and school characteristics. The regression results are presented in Table C.2. Differential attrition did not occur between the 4th grade and 5th grade groups in the process of fuzzy joining.

Table C.2. Student attrition analysis on the process of fuzzy joining the panel data (constructed from the baseline and end-line data) with the pre-baseline data

	(1)	**(2)**
(Intercept)	0.325^{***}	
	(0.046)	
Grade 4	0.009	0.076
	(0.044)	(0.065)
Female	0.043	0.031
	(0.036)	(0.033)
Z scores baseline	0.005	-0.018
	(0.022)	(0.028)
Franco-Arab school	-0.071	
	(0.047)	
Basic reading pilot school	-0.104	
	(0.068)	
School fixed effect	No	Yes
Adj. R^2	-0.001	0.368
Num. obs.	686	686

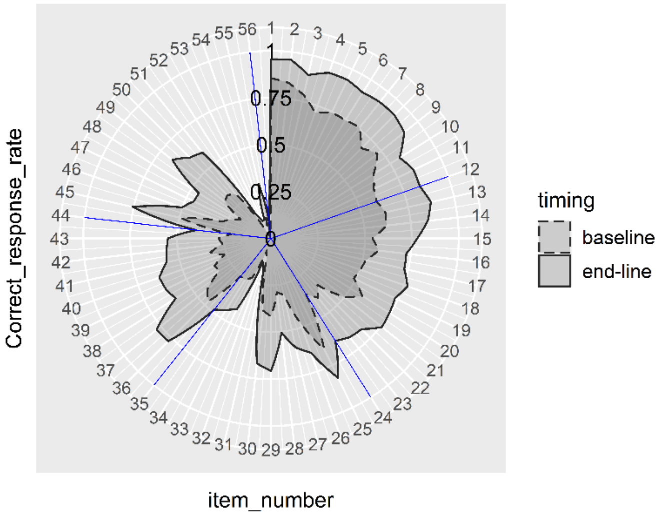
$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

(1) Robust standard errors are clustered at the school level and shown in parenthesis.
(2) All test scores are normalized by the mean and the standard deviations of the baseline scores of the 4th grade students. The maximum scale (full mark) of math test score is 44.

Graphs and table on math test results

1. Correct response rates by math items

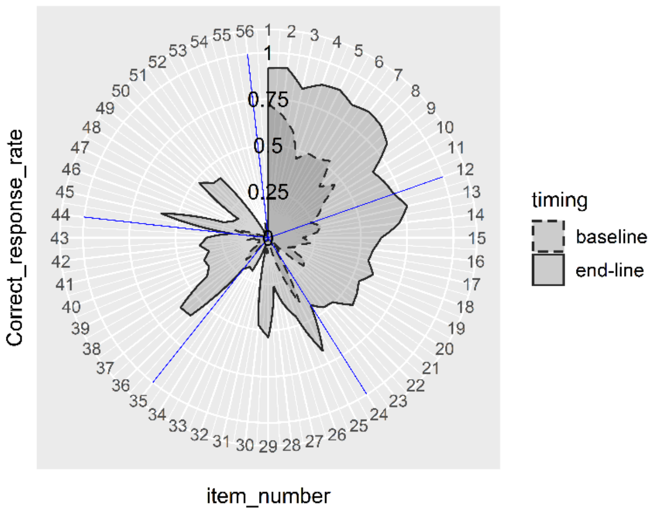
Graph D.1. Correct response rates by math item (4th grade students)



Note: The number of students in the 4th grade is 357.

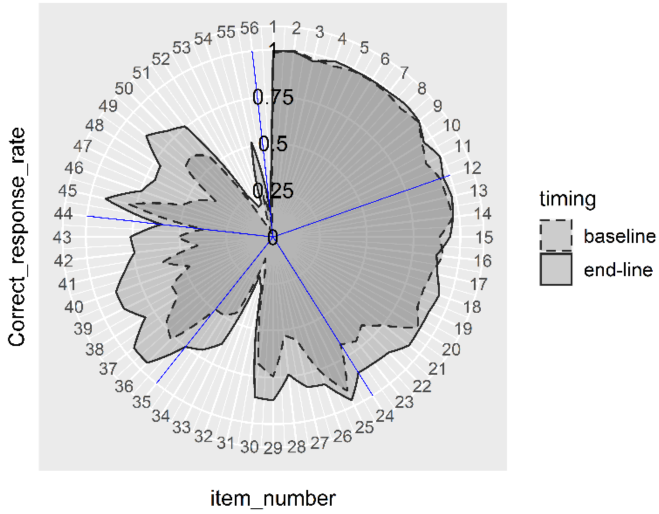
Graph D.2. Correct response rates by math item

(4th grade students below the median baseline scores)



Note: The number of 4th grade students below the median baseline scores is 177.

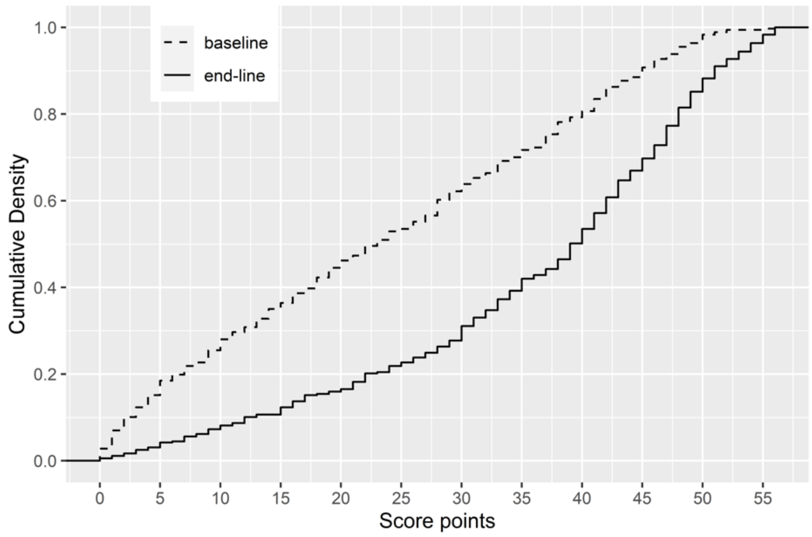
Graph D.3. Correct response rates by math item
(4th grade students over the median baseline scores)



Note: The number of 4th grade students over the median baseline scores is 180.

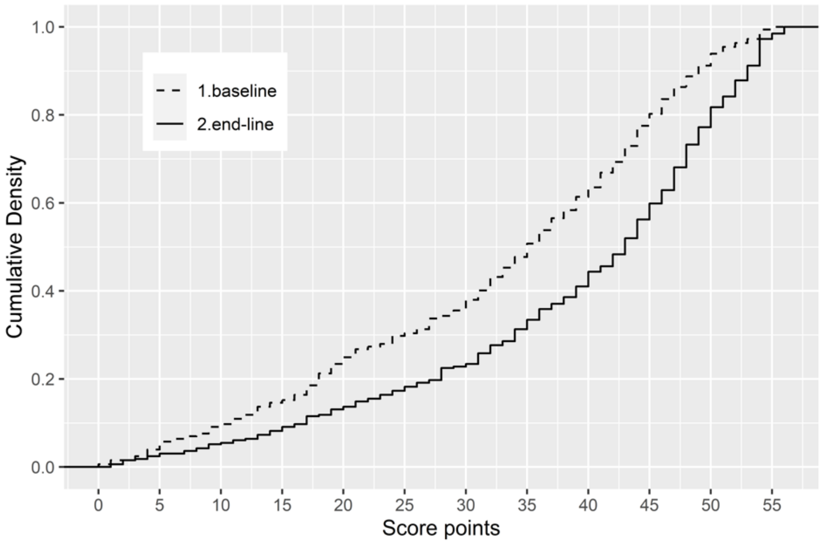
2. Cumulative density curves of test scores

Graph D.4. Cumulative density curve of test scores of 4th grade students
(baseline and end-line)



Notes (1) The number of students in the 4th grade group is 357.
(2) The maximum raw score (full marks) of the math test is 56.

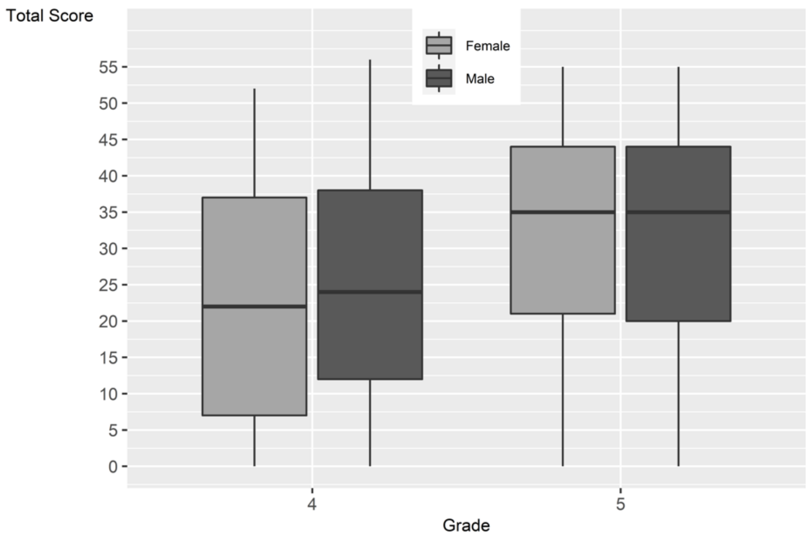
Graph D.5. Cumulative density curve of test scores of 5th grade students (baseline and end-line)



Note (1) Number of students in the 5th grade group is 329.
Note (2) Maximum scale (full marks) of the math test score is 56.

3. Boxplots of test scores by genderand by grade

Graph D.6. Boxplot of baseline test score by gender and grade (4th and 5th grade students)



Note: The number of students in the 4th grade is 357. The number of students in the 5th grade is 329.

4. Number of students by test score levels and gender

Table D.1. Number of 4th and 5th grade students by test score levels and gender

		4th grade baseline test score				5th grade baseline test score			
		Higher	Med.	Lower	Total	Higher	Med.	Lower	Total
Sex	Male	43	86	35	164	44	88	49	181
	Female	45	92	56	193	30	79	39	148
	Total	88	178	91	357	74	167	88	329

Tables on the robustness check of the estimated impact of PMAQ

Table E.1. Regression results of robustness check (1)

	(I) **4th / 5th** Pooled Model (1-b)	**(II)** **4th / 5th** Pooled Model (1-c)	**(III)** **4th / 5th** Panel: 2t Model (1-c)
(Intercept)	0.686***		
	(0.160)		
Grade 4 (Treatment) × Post (λ)	0.374***	0.389***	0.390***
	(0.092)	(0.082)	(0.077)
Grade 4 (Treatment)	-0.543***	-0.627***	-0.513***
	(0.138)	(0.102)	(0.109)
Post	0.470***	0.440***	0.411***
	(0.071)	(0.070)	(0.056)
Female	-0.174**	-0.195***	-0.141*
	(0.081)	(0.063)	(0.075)
Franco-Arab school	-0.189		
	(0.319)		
Basic reading pilot school	-0.323		
	(0.290)		
Post × Female	-0.071	-0.039	-0.037
	(0.062)	(0.041)	(0.040)
Post × Franco-Arab school	-0.018		
	(0.248)		
Post × Basic reading pilot school	-0.017		
	(0.142)		
School fixed effect (dummy)	No	Yes	Yes
Student fixed effect (dummy)	No	No	No
Adjusted R-Squared	0.150	0.508	0.514
Observations	1,995	1,995	1,372

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

(1) Robust standard errors are clustered at the school level and shown in parenthesis.

(2) The baseline and end-line data are pooled in columns (I) and (II). The panel data in column (III) is constructed from the baseline and end-line data.

(3) All test scores at each round of the survey are normalized by the mean and the standard deviations of the baseline scores of the 4th grade students. The maximum scale of test score is 56.

(4) Model (1-b) and (1-c) are expanded from model (1-a) by adding the interaction term of the end-line dummy and school or student characteristics.

Table E.2. Regression results of robustness check (2)

	(I) **4th / 5th** Panel: 2t (Sub-sample) Model (1-b)	**(II)** **4th / 5th** Panel: 2t (Sub-sample) Model (2)
Grade 4 (Treatment)×Post (λ)	0.280^{***}	0.271^{**}
	(0.078)	(0.107)
Grade 4 (Treatment)	-0.183^{**}	
	(0.075)	
Post	0.637^{***}	0.574^{***}
	(0.048)	(0.066)
Female	-0.102	
	(0.069)	
Post×Female	-0.140^{**}	
	(0.063)	
School fixed effect (dummy)	Yes	No
Student fixed effect (dummy)	No	Yes
Adjusted R-Squared	0.467	0.713
Observations	668	668

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

(1) Robust standard errors are clustered at the school level and shown in parenthesis.

(2) All test scores are standardized with the mean and standard deviation of the baseline scores of the whole sample of 4th grade students in the original panel data. The maximum scale (full marks) of math test score is 56.

(3) Panel data (baseline and end-line) is composed of students with baseline scores from 10 to 37 score points (the 2nd and 3rd quartile of 4th grade students).

(4) Model (1-b) is expanded from model (1-a) by adding the interaction term of the end-line dummy and student characteristics.

Table E.3. Regression results of robustness check (3)

	(II) **4th / 5th** Panel: 2t (Sub-sample) Standardized Model (1-a)	**(III)** **4th / 5th** Panel: 2t (Sub-sample) Standardized Model (2)
Grade 4 (Treatment) × Post (λ)	0.343^{***}	0.343^{***}
	(0.086)	(0.120)
Grade 4 (Treatment)	-0.669^{***}	
	(0.094)	
Post	0.443^{***}	0.438^{***}
	(0.067)	(0.082)
Female	-0.228^{***}	
	(0.081)	
Post × Female	-0.010	
	(0.048)	
School fixed effect (dummy)	Yes	No
Student fixed effect (dummy)	No	Yes
Adjusted R-Squared	0.490	0.811
Observations	1,036	1,036

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

(1) Robust standard errors are clustered at the school level and shown in parenthesis.
(2) All test scores are standardized with the mean and standard deviation of the baseline scores of whole samples of the 4th grade students in the panel data. The maximum scale (full mark) of math test score is 56.
(3) Panel data (baseline and end-line) is composed of all 4th grade students, and a sub-sample of 5th grade students from schools which organized additional classes for the students.

Chapter 5. External Validity of Evidence from El Salvador and Niger for Other Low- and Lower-Middle-Income Countries

Chapters 2 to 4 presented the results of impact evaluations on the package of interventions, the ESMATE program in El Salvador, and PMAQ in Niger. Both packages of interventions were found to be effective in improving math learning outcomes. The question is, are they applicable to other low- and lower-middle-income countries? This chapter first reviews the literature on the applicability of evidence to other contexts, including the analytical framework called "mechanism mapping" proposed by Williams (2020). This chapter employs the framework to explore the applicability of evidence from El Salvador and Niger to other low- and lower-middle-income countries, mainly focusing on those in sub-Saharan Africa.

5.1 Methods to examine the external validity of evidence for other countries

In the 2000s, the international community underlined the importance of scientific and credible evidence for policymaking and practices in development aid (Center for Global Development, 2006). The number of impact evaluations on basic education in developing countries has dramatically increased in response to the global awareness of the importance of evidence (Sabet and Brown 2018). Given this increase, systematic reviews were conducted to draw policy implications through the process of collecting evidence on a specific domain, examining the quality, and synthesizing them (White and Waddington, 2012). In one recent systematic review in basic education, Snilstveit et al. (2016) argued that a positive impact on student learning was consistently found in structured pedagogy programs. The program was a combination of teaching and learning materials and other supplementary interventions, such as teacher training (Snilsveit et al., 2016). McEwan (2015) conducted a meta-analysis using data from 77 different experiments on interventions for improving learning, and found that treatment with a computer or instructional technology had a positive impact on learning. Similar research was implemented by Conn (2017), using data from 66 experiments in sub-Saharan Africa, which identified a positive impact in the interventions on pedagogical techniques. Although the systematic reviews agreed that pedagogical interventions consistently had a positive impact on student learning (Evans and Popova, 2016), the policy implications from the reviews were too broad to apply to specific contexts (Williams, 2020). Across different systematic reviews, the categories in the grouping of interventions were not standardized, and the grouping of a package of different types of interventions into a specific category largely depended on the judgment of the authors of the reviews (Evans and Popova, 2016).

Given the methodological limitation of systematic reviews for policymaking and practice, Williams (2020) proposed a method called "mechanism mapping" to explore the applicability of evidence to other contexts. The method consists of several steps, starting from laying out a theory of change in the intervention. A theory of change is a causal chain from initial inputs to the intended final outcomes via activities, outputs, and intermediate outcomes of the intervention. The second step is to raise important contextual assumptions related to each step in the causal chain. The contextual assumptions include social, economic, and political features on the background of the experiment, characteristics of target groups of intervention, competence, and resources of the organization that implements intervention (Williams, 2020). The final step is laying out the actual contextual situations corresponding to the assumptions. Differences between the contextual assumptions and actual situations suggest that the causal chain in the theory of change might be interrupted. Policymakers and practitioners can identify the stage at which the causal chain is likely to be interrupted, and consider countermeasures. Whereas mechanism mapping cannot perfectly predict the external validity of evidence, the method makes a theory of change, contextual assumptions, and actual contextual characteristics explicit, which helps policymakers and practitioners to explore the applicability of evidence in a structured manner (Williams, 2020).

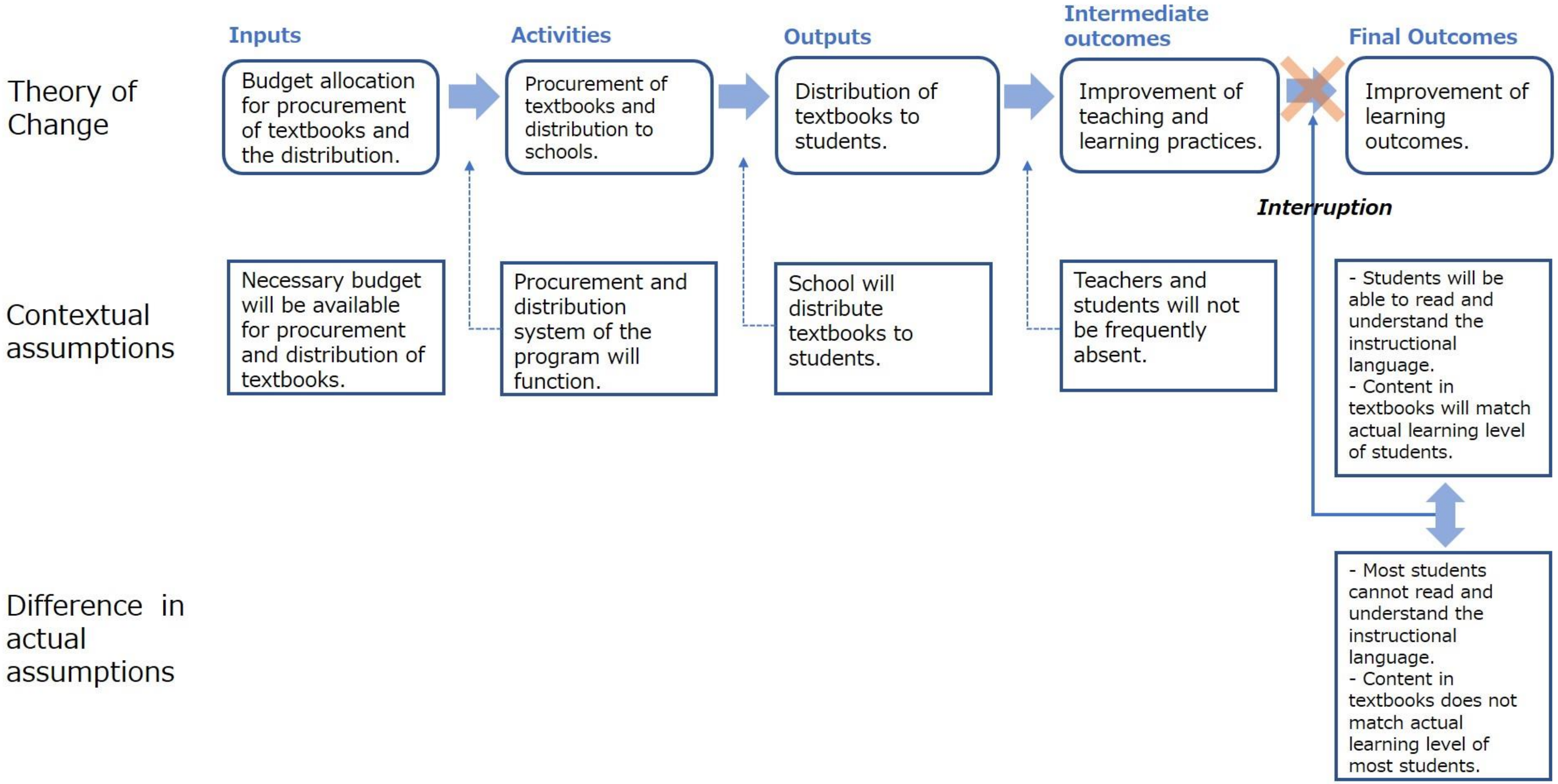
Mechanism mapping is similar to the "generalizability framework" proposed by Bates and Glennerster (2017). The generalizability framework begins by laying out a theory of change and the necessary local conditions for the theory. The subsequent steps are to check the related evidence that supports the theory of change, and to assess the feasibility of successful implementation of the intervention. Although both methods are similar in focusing on the mechanism of intervention to explore the external validity, mechanism mapping is more practical because it explicitly captures the difference between the contextual assumptions and actual situations. The method of mechanism mapping supposes a specific country in principle, but it can be employed to explore the external validity of evidence more broadly. This chapter utilizes mechanism mappings to examine the applicability of the ESMATE program and PMAQ to other low- and lower-middle-income countries, mainly those in sub-Saharan Africa, considering possible differences between the contextual assumptions in a theory of change and situations in the countries.

5.2 Applicability of the ESMATE program to other low- and lower-middle-income countries

Before doing the mechanism mapping of the ESMATE program, this section starts with the exercise on a simple example of the textbook distribution program, called "School Assistance Program," in Kenya, which was reported by Glewwe et al. (2009). The program in Kenya, funded by a Dutch non-profit organization, distributed math and English textbooks to grade 3, 5 and 7 students in 25 primary schools in the Busia and Teso districts.[64] Figure 11 depicts the mechanism mapping of the textbook distribution program.

[64] The program funded by Dutch non-profit organization was called "School Assistance Program." For grade 4 and 6, only English textbooks were distributed. For grade 8 students, only science textbooks were delivered. The experiment continued for two years. For grade 4 and 6 in year 2 of the research, math textbooks were distributed.

Figure 11 Mechanism mapping of "School Assistance Program" in Kenya



The core problem addressed in the program was the shortage of textbooks for primary students in the targeted area. Before the program started, around 80 percent of students in the sampled schools were in the classroom with less than one math textbook for every 20 students (Glewwe et al., 2009). The program tried to increase the textbook to student ratio by distributing textbooks to students, thereby improving student learning outcomes (Figure 11). Whereas the program effectively increased the textbook/student ratio, the average impact on student learning was close to zero and not statistically significant (Glewwe et al., 2009). The program had a positive impact only for students with the highest quintile of baseline test scores, and did not affect the pedagogy of teachers (Glewwe et al., 2009). Since the distributed textbooks were written in English, a third language for most students in the country, Glewwe et al. (2009) argued that the mismatch between the student learning level and the content of the textbook was a major factor behind the heterogeneous impact. In Figure 11, the contextual assumptions in the theory of change, "students will be able to read and understand the instructional language in textbooks" and "contents in textbooks will match actual learning level of students," did not hold for most students, which brought the unintended heterogeneous impact on learning outcomes.

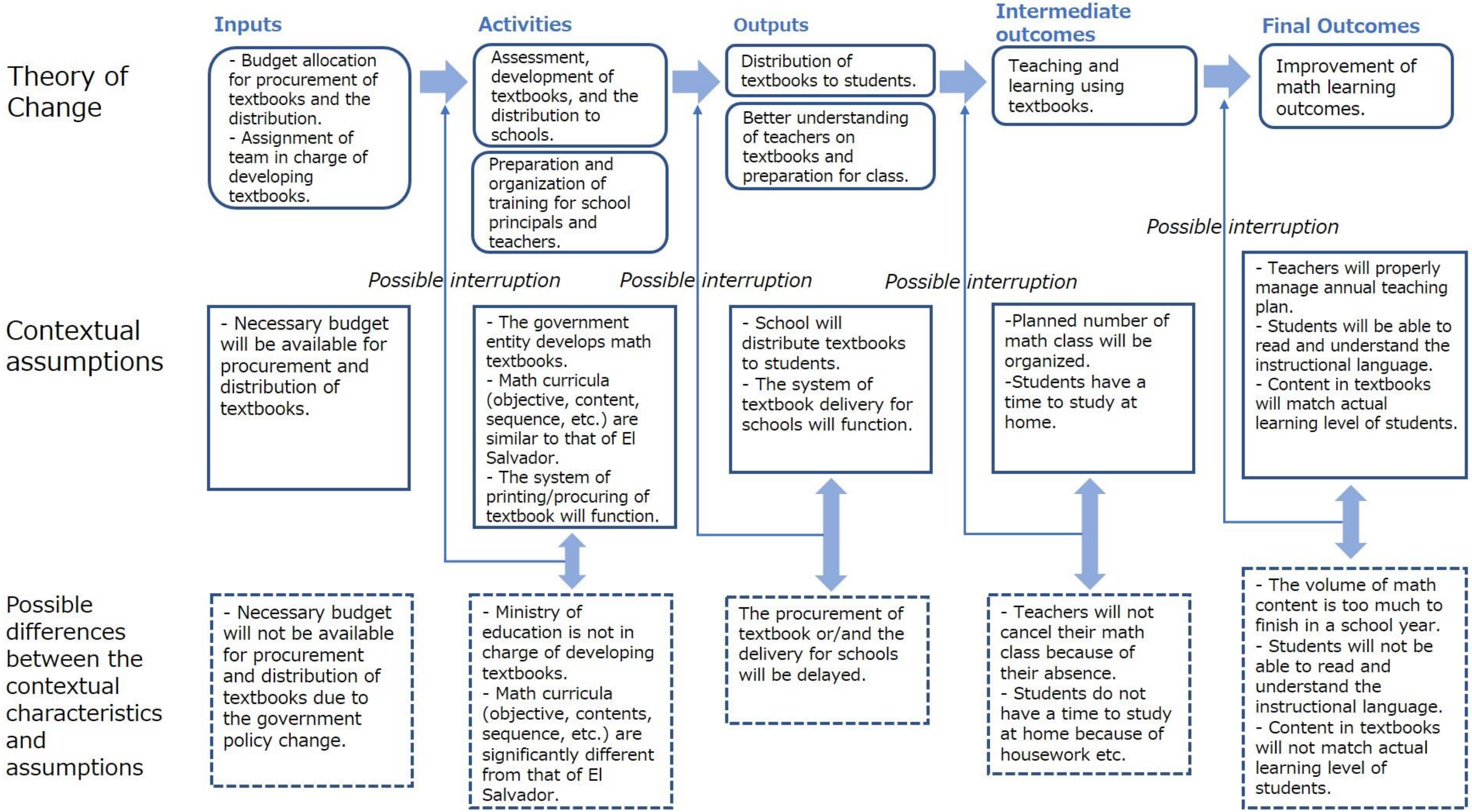
As reported in Chapters 2 and 3 of this dissertation, with technical cooperation from JICA, the Ministry of Education in El Salvador developed learner-friendly math textbooks called the ESMATE textbooks. The project conducted a student math assessment in the country before textbook development. Based on the findings in the assessment, the subject contents were subdivided and sequenced in the textbook to ensure small-step learning by students. The page in the ESMATE textbook was structured in four steps: (a) present the theme of the lesson, (b) pose example problems, (c) explain the general principle, and (d) provide exercises. Through this process, the ministry tried to avoid the mismatch between the learning content and actual learning level of students. The ministry also developed teacher guides and student workbooks as supplementary teaching and learning materials. The Ministry of Education distributed a set of teaching and learning materials to schools, and provided introductory training for teachers and school principals. Teachers regularly reviewed their plans and practices, referring to the math test results in their classes.

Figure 12 presents the mechanism mapping of the ESMATE program applied in other low- and lower-middle-income countries, which includes a theory of change and contextual assumptions of the program, and the possible differences between the contextual assumptions and situations in other low- and lower-middle income countries. The mechanism mapping supposes the program to be implemented by the Ministry of Education in the countries. First, the ministry organizes or reviews the student assessment results and math curricula in the country, and then develops the plan of textbooks, referring to the ESMATE textbooks. If the learning contents are not systematically arranged in the math curricula in the country, it would be necessary to revise it before textbook development.

The development of math textbooks depends on the contextual assumptions in different aspects, for example, at institutional aspect like math curricula, cultural aspect like mother tongue and instructional language, and organizational aspect such as the capacity for printing/procuring and delivering textbooks to schools. Since the ESMATE program primarily targets the development of math textbooks, there are a number of important contextual assumptions that support the theory of change as shown in Figure 12. When the program is applied to other developing countries, there are differences between the assumptions and actual situations.

First, the system of textbook development varies among countries. In some countries, textbooks for basic education are developed by governmental agencies, while in others, the role of the government is just to review and adopt textbooks developed by private companies. Other countries employ a mixed arrangement of the two (Japan Textbook Research Center, 2020). For countries where the government reviews and adopts a math textbook developed by private companies, it is necessary to modify the activities in the theory of change. For example, revision of standards of the government in review and adoption of textbooks would be one of the activities to be added to the causal chain of the program.

Figure 12. Mechanism mapping of the ESMATE program applied to other low- or lower-middle-income countries



Second, the content of math curricula varies among countries. The curricula define the learning objectives and outline learning contents and methods (UNESCO-IBE, 2013), which are the foundation of textbook development. In math curricula of basic education, El Salvador aims at developing competencies of mathematical and logical reasoning, communication using mathematical language, and applying mathematics in daily life (Currículo al Servicio del Aprendizaje). The curriculum emphasizes the importance of the learning process where students participate in activities in class, work on problem-solving (Currículo al Servicio del Aprendizaje). For countries that intend different types of competencies in math curricula, it is necessary to carefully examine whether and how the ESMATE textbook might be applicable.

Third, textbooks are written in the instructional language in the country, but the mother tongue for most students can be different in low- and lower-middle-income countries. In El Salvador, the instructional language is Spanish, which is also the mother tongue for most students. The linguistic diversity of the country is very low. According to the ranking of the linguistic diversity index (UNESCO, 2009), which is the probability that any two randomly selected people would have a different mother tongue, El Salvador is situated in the lowest quintile of the countries in the world (196th out of 202 countries that have available data). The likelihood of difference between the mother tongue and the instructional language increases in countries with higher linguistic diversity. Bashir et al. (2018) cited linguistic diversity as one of the challenges for educational development in sub-Saharan Africa. For the countries where instructional language and mother tongue are different, it is essential for the ministry to consider measures to mitigate the language barrier for students, especially in the early grades.

Lastly, the capacity for printing/procuring and delivering textbooks to schools could be a bottleneck for applying the ESMATE program for other low- and lower-middle-income countries. The delay or loss in the delivery of textbooks to schools disrupts math teaching plans at schools. In El Salvador, the ministry regularly distributed a package of stationaries to students at the beginning of every school year. The logistical capacity realized the procurement and distribution of ESMATE textbooks nationwide (JICA, 2019). On the other hand, in sub-Saharan Africa the distribution of textbooks was very inefficient including large losses in the distribution process and delay (Read, 2015). For countries with limited capacity for printing/procuring and delivery of textbooks to schools, it would be essential for the ministry to identify a bottleneck in the system and consider how its functionality could be improved. The measures to improve functionality differ among countries, depending on the procurement and delivery systems of textbooks.

Whereas the ESMATE program included various types of interventions such as teacher training and mutual review meeting of teachers to improve learning, the challenges that the program addressed were limited to pedagogical aspects. The other challenges than pedagogical ones might hamper the causal chain in the ESMATE program. For example, the ESMATE textbook was designed to be finished with 160 hours of math class in a year (JICA, 2019).[65] In Latin America, teacher absenteeism is a profound issue in education (Bruns and Luque, 2015). Due to teacher absenteeism, the number of math class hours might be short of covering the textbook well. Even in the case, teachers might not change their teaching plan by sticking to finish the prescribed syllabus. The ESMATE program also tried to strengthen student study at home by regular homework assignment; however, students might not be able to have their time to study at home due to various reasons including housework. While the program included an intervention to sensitize the representative of parent association, it would be necessary to strengthen the intervention to better involve parents to support their children at home when the ESMATE program be applied to other countries.

5.3 Applicability of PMAQ to other low- and lower-middle-income countries

As mentioned in Chapter 4 of this dissertation, with technical cooperation from JICA, the Ministry of Education in Niger organized a set of training on the democratic establishment of school management committees (SMCs) and school action plan development cycle for all public primary schools in the 2007-08 and the 2008-09 school years (Honda and Kato, 2013). The training intends to

---

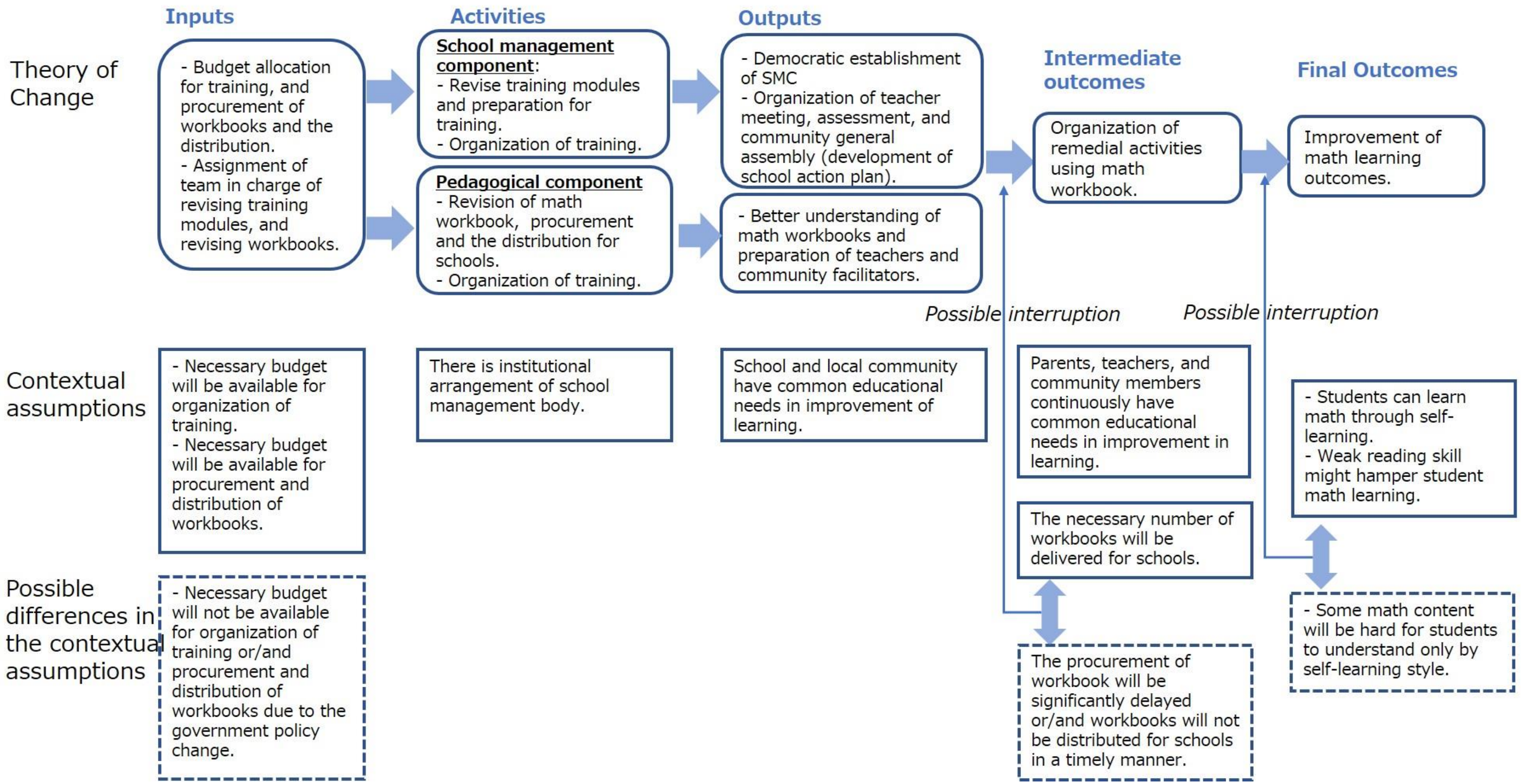
[65] The number of hours of math class in math curricular was 200 for grades 2 through 9 in El Salvador.

establish SMCs with leadership through a secret ballot and enhance their capacity to lead the process of the school action plan cycle with community participation (Kunieda et al., 2020). The democratically elected accountant of SMC manages the resources mobilized from the local community and shares the accounting report regularly at the community general assembly. In 2018, PMAQ was scaled up in the Tillaberi region based upon the foundation of functioning SMCs. When the applicability of PMAQ to other low- and lower-middle-income countries is considered, it is thus appropriate to think of PMAQ as well as the training on the democratic establishment of SMCs and school action plan development cycle as a set of interventions.

A systematic review of evidence on school-based management in developing countries argued that one of the challenges in the type of intervention was the elite capture of SMCs (Carr-Hill et al., 2018). The process from democratic establishment of SMCs to regular information-sharing on the mobilized resources was designed to prevent the decision-making be monopolized by a limited number of SMC members or school principals. The other challenge posed by Carr-Hill et al. (2018) was low level of capacity within community. Communities with limited educational attainment including high level of illiteracy could not plan and implement relevant activities to improve learning (Beasley and Huillery, 2016). In PMAQ, the process of developing school action plan including remedial activities was designed to address the low level of educational attainment in the local community (Hara et al., 2020). The assessment results are visually presented in a manner that local community could easily understand the problem. The SMC representatives facilitate the discussion at community general assembly involving parents, teachers, and community members.

Figure 13 describes a theory of change of PMAQ including the training on the democratic establishment of SMCs and school action plan development cycle. The Ministry of Education prepares a set of training modules for the school principal and members of SMC, adapting the modules to the institutional arrangement (e.g., member composition) of the school management body in the country. Then, the ministry provides a series of training for the school principal and members of the SMC. After the first training, the school principal organizes the community general assembly, where the SMC is democratically established. After a second training for SMC members, the school principal holds a teacher meeting, and then organizes a math assessment for all students in the school. The SMC organizes a community general assembly to share the summary of assessment results, and facilitates discussions among parents, teachers, and community members. The Ministry of Education also adapts the math workbooks on numbers and the four basic operations in the local context of the country. The Ministry organizes a training for teachers and community facilitators on the content and use of math workbooks.

Figure 13. Mechanism mapping of PMAQ applied to other low- or lower-middle-income countries



The two causal chains in Figure 13 are joined at the intermediate outcome stage, that is, the organization of remedial activities using math workbooks. The two components are complementary in that each component addresses the contextual assumption of the other. For example, if teachers or students were frequently absent, distributed math workbooks would not be effectively utilized. In sub-Saharan Africa, teacher absenteeism was a profound problem that caused a shortage of learning time at schools (Bold et al., 2017; World Bank, 2017). Similarly, the impact of remedial activities organized by the collaboration of parents, teachers and community members on student math learning diminishes, unless teachers have sufficient pedagogical skills and content knowledge. The pedagogical skills and content knowledge of teachers were surprisingly low in sub-Saharan Africa (Bold et al., 2017). For example, around 70 percent of teachers in Nigeria did not have knowledge of the language and math of primary grade 4 curricula. The pedagogical component of PMAQ, composed of the distribution of math workbooks and training, complements the pedagogical skills and content knowledge of teachers. The first component of PMAQ covers the contextual assumption required for the second component of the package by realizing extra-curricular remedial activities. The second component of PMAQ also similarly addresses the contextual assumption required for the first component of the package by providing a set of learning materials and training.

The causal chain of school management component in PMAQ supposes the existence of an institutional arrangement of a school management body that involves the local community. Since the mid-1990s, the school management body, which took different names, such as the school management committee, school council, or school board, was widely introduced in developing countries under the school-based management policy (Barrera-Osorio et al., 2009; Bruns et al., 2011). The principal idea of school-based management was to decentralize authority and resources at the local level, handing over decision-making to principals, teachers, parents, and sometimes to students or other community members, with the aim of improving the efficiency of financing and quality of education services (Barrera-Osorio et al., 2009). Several kinds of authorities or resources can be devolved to the school management body, including budget allocations, hiring and firing of teachers and other school staff, procurement of textbooks and other educational material, infrastructure improvement, monitoring and evaluation of teacher performance and student learning outcomes (Bruns et al., 2011). Whereas PMAQ utilized the school management body as an institutional arrangement to involve parents and community members, the package did not suppose particular authorities or resources devolved from the central government for SMCs. Thus, the contextual assumption on the institutional arrangement of the school management body is likely to hold in low- and lower-middle-income countries.

Second, PMAQ supposes that the school and local community potentially have common educational needs in the improvement of learning. In the PMAQ model, SMC organizes student assessment in math for all grades and presents the results at the community general assembly. Students should obtain full marks on the math assessment, as the math items are from the curricula that they have learned. The gap between the full mark and the actual level of correct response rate highlights the problem in math learning, which leads the discussion at the community general assembly to develop school action plan. In sub-Saharan Africa, over 80 percent of children of primary and lower secondary school age were not mastering the basics of reading and mathematics (UNESCO, 2017). A large volume of children in sub-Saharan Africa was not mastering even the basics of mathematics, including numbers and the four basic operations (PASEC, 2016). The situation of the global learning crisis has worsened by the long-term closure of schools due to the COVID-19 pandemic, especially in low- and lower-middle-income countries (UNESCO, 2021). In the situation of low-level learning in developing countries, parents, teachers, and community members are likely to identify common educational needs in improving math learning through information-sharing of student assessment results.

Contrary to the two contextual assumptions on school management component, the capacity for printing/procuring and delivering workbooks to schools could be a bottleneck for the package, similar to the ESMATE program. In the case of scaling up PMAQ in Niger, the distribution of math workbooks was postponed for one school year because of the delay in procurement. Thus, for countries with limited capacity for printing/procuring and delivery of learning materials to schools, it would be necessary for the ministry to develop a plan of scaling-up PMAQ carefully, which could be done by several phases (geographically dividing the country into several phases), or targeting only disadvantaged schools.

5.4 Measures to enhance the applicability of PMAQ and the ESMATE program

As discussed in section 5.2, the ESMATE program as a whole is not easily applicable to other low-and lower-middle-income countries, since the major component of the program is textbook development, which requires several contextual assumptions. Whereas the ESMATE program as a whole is not easily applicable to other countries, textbooks remain important intermediates that link curricula, teachers and students. Textbooks also cover math curricula. Teachers can improve their pedagogical skills using quality textbooks. For the countries where the ESMATE program is applied, the country should enhance its impact by addressing some contextual assumptions. For example, in countries where the mother tongue of students and the instructional language are different, it is necessary to mitigate the language barrier, especially for early graders. Otherwise, the heterogeneous impact by the initial learning levels, similar to the experiment in Kenya, might happen.

The applicability of PMAQ can be enhanced by addressing the contextual assumptions and upgrading the package. First, in the PMAQ model, student workbooks were designed so that students could learn numbers and four basic operations by self-learning style. While the remedial activities with the self-learning style improved math learning in Niger, there were math contents that were difficult for students to learn well. For example, primary 4th grade students had a difficulty in solving the items of two-digit subtraction with borrowing, three to four-digit addition with carrying, and subtraction with borrowing. It might be possible to overcome the limitation of PMAQ by incorporating lesson-style remedial activities utilizing the teaching and learning materials of the ESMATE program. Selecting several key units or sub-units from the teaching and learning materials, and integrating supplementary math lessons using the materials in the remedial activities would be an option to improve the impact of PMAQ.

This chapter explored the applicability of the ESMATE program and PMAQ to other low- and lower-middle-income countries, using the analysis framework called “mechanism mapping” proposed by Williams (2020). While the two packages of interventions shared the overall goal of improving student math learning, the approaches were different but complementary. For example, whereas the former covered the math curricula in basic education, the latter focused on a specific domain in math, that is, numbers and four basic operations, which are foundational skills for students to learn math in basic education. The ESMATE program improved student math learning on average, but the difference in math learning levels became tangible in a school year, as discussed in Chapter 3 of this dissertation. It causes the mismatch between the content of math textbook and student learning level, which might hamper the causal chain in the ESMATE program (Figure 13). In this aspect, PMAQ can be complementary to the ESMATE program since remedial activities in the PMAQ model helps students with lower math achievement catch up. The process of information-sharing and discussions among local stakeholders clarifies the status of learning levels and facilitates teachers to provide students with additional guidance and instructions that match their learning levels. The synergy of the two packages is worth exploring for the improvement of math learning in low- and lower-middle-income countries.

Appendix 18. Utilizing Evidence to Improve Educational Aid: A Case study of the “Read India Program” by the Indian NGO “Pratham”

Reference

Barrera-Osorio, Felipe, Tazeen Fasih, Harry Anthony Patrinos, Lucrecia Santibáñez. 2009. *Decentralized Decision-making in Schools: The Theory and Evidence on School-based Management*. Washington D.C.: World Bank.

Bates, Marry Ann and Rachel Glennerster. 2017. “The Generalizability Puzzle.” *Stanford Social Innovation Review*, Summer 2017.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jacob

Svensson, and Waly Wane. 2017. “Enrollment without Learning: Teacher Effort, Knowledge and Skill in Primary Schools in Africa.” *Journal of Economic Perspectives*, November 2017.

Bruns, Barbara, Deon Filmer, and Harry Partrinos. 2011. *Making Schools Work: New Evidence on Accountability Reform*. Washington D.C.: World Bank.

Center for Global Development. 2006. *When will we ever learn? Improving lives through impact evaluation.* Washington D.C.: Center for Global Development.

Conn, Katharine M. 2017. “Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations.” *Review of Educational Research*, Vol. 87 (5): 863–898.

Evans, David K., and Anna 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.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. “Many Children left behind? Textbooks and Test Scores in Kenya.” *American Economic Journal: Applied Economics* 1 (1): 112-35.

JICA 2019. *Project Completion Report of Project for the Improvement of Mathematics Teaching in Primary and Secondary Education in El Salvador. (in Japanese).*

Japan Textbook Research Center. 2020. *Kaigai kyokasho seido tyousa kenkyu hokokusho (Report of the institutions on textbooks overaseas).* Tokyo: Japan Textbook Research Center. (in Japanese).

Maruyama, Takao, Kengo Igei and Seiichi Kurokawa. 2021. “Community-wide support for primary students to improve learning: Empirical evidence from Madagascar.” *RISE Annual Conference 2021*. Retrieved from https://riseprogramme.org/sites/default/files/inline-files/Maruyama_Community_wide_support_primary_students_improve_learning_Empirical_evidence_Madagascar.pdf

McEwan, Patrick J. 2015. “Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments.” *Review of Educational Research*, Vol. 85 (3): 353–394.

The ministry of education in El Salvador. 2008. *Currículo al servicio del aprendizaje*. Retrieved from https://www.mined.gob.sv/download/curriculo-al-servicio-del-aprendizaje/

PASEC 2016. *PASEC 2014 – Performances du système éducatif nigérien: Compétences et facteurs de réussite au primaire*. Dakar: PASEC. CONFEMEN.

Read, Tony. 2015. *Where Have All the Textbooks Gone? Toward Sustainable Provision of Teaching and Learning Materials in Sub-Saharan Africa*. Washington D.C.: The World Bank.

Sabet, Shayda Mae and Annette N. Brown. 2018. "Is impact evaluation still on the rise? The new trends in 2010–2015." *Journal of Development Effectiveness*, 10:3: 291-304.

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt and Emmanuel Jimenez. 2016. *The Impact of education programmes on learning and school participation in low- and middle- income countries*. London: International Initiative for Impact Evaluation.

UNESCO 2009. *Investing in Cultural Diversity and Intercultural Dialogue*. Paris: UNESCO.

UNESCO 2017. "More than one-half of children and adolescents are not learning worldwide." *Fact Sheet No. 46* (UIS/FS/2017/ED/46).

UNESCO 2021. *Pandemic-related disruptions to schooling and impacts on learning proficiency indicators: A focus on the early grades*. Montreal: UNESCO Institute for Statistics.

UNESCO International Bureau of Education 2021. "Curriculum." Retrieved from http://www.ibe.unesco.org/en/geqaf/core-resources/curriculum

White, Howard and Hugh Waddington. 2012. "Why do we care about evidence synthesis? An introduction to the special issue on systematic reviews." *Journal of Development Effectiveness*, 4 (3): 351-358

Williams, Martin, J. 2020. "External Validity and Policy Adaptation: From Impact Evaluation to Policy Design." *The World Bank Research Observer*, Vol. 35 (2): 158–191.

Appendix 18. Utilizing Evidence to Improve Educational Aid: A Case study of the "Read India Program" by the Indian NGO "Pratham" [66]

Chapter 5 of this dissertation explored the applicability of the evidence from El Salvador and Niger for other low- or lower-middle-income countries using mechanism mapping. The analysis framework shed light on the contextual assumptions to be reviewed in applying evidence to other contexts, and strengthened a theory of change in intervention by demonstrating the potential gaps between contextual assumptions and actual situations. While evidence and the examination of external validity assist development agencies to improve educational aid, the success largely depends on how the agencies actively utilize the evidence in their strategy development and operation management. This chapter presents a case study of the Indian NGO "Pratham," a prominent example of an organization that improved and expanded the operation through a series of impact evaluations (Banerjee et al., 2017). Through the case study, this chapter also presents lessons learned for other development agencies.

6.1 Approaches to improving evidence use in practice contexts

In response to the global awareness on the importance of scientific evidence for development aid in the 2000s, the number of impact evaluations, including those in basic education, has significantly increased (Sabet and Brown, 2018). However, the results of impact evaluations were not fed back well into the strategies and practices of development agencies (Dhliwal and Tulloch, 2012; Langer et al., 2015; Shah et al., 2015). Langer et al. (2015) claimed that neither policymakers nor practitioners were commonly incentivized to use research evidence. This is because a strict process of quality control, such as for drugs in healthcare, was not required for a program in international development. Shah et al. (2015) argued that impact evaluations were often "researcher-driven," as researchers focused on building the global body of knowledge on "what works"; however they did not pay much attention to

---

[66] This work was supported by JSPS KAKENHI Grant Number JP20K02559.

operational issues.

In a broad sense, enhancing evidence-use in development aid is a question of improving research use in practice contexts. Nutley et al. (2007) have proposed three approaches to improving research use in practice in the social sector: (i) research-based practice model, (ii) embedded research model, and (iii) organizational excellence model. In the first model, research use is the responsibility of individual practitioners. Individual practitioner reviews the existing research and integrate the findings with their own knowledge to manage daily practices and decision-making. In the second model, individual practitioners seldom directly access the findings of individual research; instead, policymakers or organization managers translate them into standards, frameworks, guidelines, procedures, and tools from findings in researches. Individual practitioners refer to the research-based guidance and tools.

The two models above suppose a linear process of research use by practitioners in which researchers provide evidence, and practitioners refer to it. Conversely, in the third model, the organizational excellence model, practitioners and researchers develop locally adapted practice through research including local experimentation in which they learn from each other. The process is interactive and iterative rather than linear as in the other two models. The three approaches proposed by Nutley et al. (2007) are not exclusive to each other but shed light on different aspects of evidence use by organizations. For example, after local experimentation, it is necessary for an organization to develop standards or guidelines to share the developed practice among their staff. The organization staff should then regularly review their practices, and check their performance with other organizations by referring to research.

The process of improving student learning in developing countries involves multiple actors, including parents, teachers, school principals, and local education officers. (World Bank, 2018). The characteristics of the actors and socio-economic contexts of students vary considerably among countries. Therefore, local experimentation helps the development agencies to test and improve the programs, as suggested in the organizational excellence model. However, using research successfully to develop and improve program depends on the leadership and management of the organization (Nutley et al., 2007). In the case of rigorous evaluation of a program, development agencies should actively interact with researchers to develop a evaluation plan, and feedback the results of the evaluation to their operation and decision-making. However, the literature on evidence-use in international development mainly supposes one-way communication from researchers to development agencies, rather than interactive process between the two entities. It does not also focus on organizational aspects such as the initiative and management of development agencies.

In the field of educational development, a prominent example of the "organizational excellence model" is the Indian NGO "Pratham" (Banerjee et al., 2017; Duflo, 2020; Dutt and Robinson, 2016). Pratham was established in Mumbai in 1994 with a mission to help children go to school and learn well. The leading program, "Read India," aims to help children learn basic reading and math. The Read India program began in 2007 and expanded throughout the country. The number of children who benefitted in the program in 2018 was estimated at around 16.6 million in 21 states (Pratham Education Foundation, 2019). Since the early 2000s, Pratham conducted a series of experiments in collaboration with researchers in development economics to verify the impact of the pedagogy employed in the program and develop scaling-up strategy (Banerjee et al., 2017; J-PAL, 2017).[67] As the case of evidence use of Pratham is longitudinal and includes several examples of impact evaluations, the case will reveal the initiative and management of the organization in utilizing evidence to improve the operations. This chapter conducts a case study of evidence use by Pratham in the Read India program, and also presents lessons learned for other development agencies.

6.2 Proof of concept of Teaching at the Right Level: experiment in Mumbai (2001–2003)

Pratham developed a pedagogy called "Teaching at the Right Level (TaRL)," and employed

[67] The impact evaluations are (i) "Can Informational Campaigns Raise Awareness and Local Participation in Primary Education in India?"; (ii) "Read India: Helping Primary School Students in India Acquire Basic Reading and Math Skills"; (iii) "Using Learning Camps to Improve Basic Learning Outcomes of Primary School Children in India"; (iv) "Improving Learning Outcomes through the Government School System in India" (J-PAL, 2017).

it in the Read India program.[68] The pedagogy was designed for children from primary grades 3 to 5 in India to support them in rapidly improving basic reading and math (numbers and four basic operations) (Banerji et al., 2005; Pratham, 2020). The pedagogy was first developed in late 2002, and was continuously refined by Pratham. The first step of TaRL is simple assessments of basic reading and math. Children are grouped according to their basic reading and math proficiency levels regardless of the grade. They learn basic reading and math through different kinds of activities using simple tools (Pratham, 2020). One of the tools for learning basic reading is "barahkhadi" chart which is a grid with consonant sounds in rows and vowels in columns. The chart is utilized to familiarize children with the correspondence between sounds and symbols. The chart is then used for children to learn syllables made up of consonants and vowels. Children can also refer to the chart when practicing writing words (Banerji and Chavan, 2016). Simple assessments of TaRL are regularly conducted, and children are re-grouped according to their assessment results.

From 2001 to 2003, Pratham conducted the first experiment on the original pedagogy in Mumbai. Volunteers, called "Balsakhi," were randomly assigned for the treatment group, and they organized remedial activities in basic reading and math. The study confirmed a positive impact on learning, which served as proof of concept of the pedagogy (Banerjee et al., 2017; Pratham, 2020).

6.3 Scaling-up TaRL by mobilizing community volunteers: experiment in Jaunpur (2005)

In 2002, Pratham started expanding the operation from Mumbai and other cities to rural areas in India (Banerji, 2020). In this context, the second experiment on TaRL was planned in Jaunpur, Uttar Pradesh, to explore the external validity of the evidence from Mumbai. The experiment by Pratham was also a "search" for scaling up strategy for TaRL in rural areas in India. A "search" is a concept proposed in the literature on bounded rationality (Simon, 1979). As the cognitive capability of any organization

[68] The pedagogy in basic reading was initially called "Learning to Read (L2R)." The pedagogy is also called "CAMaL (Combined Activities for Maximized Learning)." The phrase TaRL has been used since around 2012 (Pratham, 2020). In this chapter, I consistently use "TaRL" to call the pedagogy.

is limited, the organization does not have sufficient information on all possible options and their potential outcomes in advance; therefore, under bounded information, an organization searches for options to achieve better results (Simon, 1979).

Pratham leaders recognized that a large proportion of children in rural areas in India had not mastered foundational skills, basic reading and math (Banerji, 2005; Duflo, 2020). The agenda for Pratham at the time was how to scale up TaRL so that it could benefit a large number of children in rural areas (Banerji, 2020). The organization took a strategy to scale up TaRL by mobilizing community volunteers (Dutt and Robinson, 2016; Banerji, 2019). While the results of the experiment in Mumbai were positive, it was not evident for Pratham how their program would work in rural areas (Banerji, 2019). The context in rural areas was significantly different from Mumbai, and the implementers of TaRL in rural areas were community volunteers. In a less structured environment than Mumbai, could TaRL be effectively implemented by community volunteers? The research agenda motivated Pratham and researchers to plan the experiment. Among the areas where Pratham was not so active, Jaunpur was selected for the experiment site, based on educational indicators, including learning outcomes (Banerji, 2019).

Pratham and researchers reviewed different literature as well as the past experiences and data of the organization to design the contents, targets, and implementation structures of interventions (Glennerster and Takavarasha, 2013). Pratham tried to develop interventions that could be replicable on a large scale, easily implemented without highly educated human resources, and were low-cost (Glennerster and Takavarasha, 2013). Pratham and researchers designed three types of interventions: (i) provision of information on village education committees; (ii) intervention (i) plus creating village-based scorecards on reading skills; and (iii) intervention (i) and (ii) plus demonstrating TaRL (Banerjee et al., 2010). Intervention (i) targeted villagers, and Pratham staff supported the facilitation of the discussions at the village meeting. In intervention (ii), Pratham staff trained the community volunteers on how to evaluate children's basic reading. In intervention (iii), Pratham staff demonstrated remedial activities in basic reading, and trained community volunteers in the methods. In the process of intervention design, Pratham and researchers laid out a theory of change for the interventions, and developed research questions. The research questions included whether the information provided can fuel community action to improve basic reading; if yes, what kind of information it was, and whether it was necessary to present a convincing solution to improve learning (Banerjee et al., 2010; Banerji, 2019).

The results of the experiment in Jaunpur revealed the impact of TaRL in a rural context as well as future challenges. While it was found that the first two types of interventions did not work, the third intervention was effective in mobilizing community volunteers and improving learning (Banerjee et al., 2010). The results confirmed that community volunteers could effectively implement TaRL to improve student basic reading in the context of rural areas. It encouraged Pratham to scale up TaRL by mobilizing community volunteers in the Read India program (Glennerster and Takavarasha, 2013). Furthermore, Pratham and researchers noted several challenges from the experiment. For example, the pace and durability of remedial activities considerably varied among community volunteers and communities (Banerji, 2020). Substantial attrition of the community volunteer was also observed, and many classes ended prematurely (Duflo, 2020). Since the intervention targeted children outside of school, the take-up was not universal (Duflo, 2020).

6.4 Scaling-up TaRL in government schools: experiment in Bihar and Uttarakhand (2008−2009), and in Haryana (2012−2013)

In the experiment in Jaunpur, the research agenda for Pratham was "how to scale up TaRL so that it could benefit a large number of children in rural areas." After the experiment on community volunteer mobilization, Pratham turned its attention to government schools, and refined the research agenda to "how it was possible to insert TaRL into a large poorly functioning government school system," which led them to plan the following experiment (Banerji, 2019; 2020). The areas in Bihar and Uttarakhand were selected because of the contrasting characteristics of initial learning levels, socio-political contexts and administrative capabilities (Banerjee et al., 2016). Pratham and researchers discussed what intervention for the government schools would bring the improvement of learning, and developed four types of interventions: (i) the distribution of teaching and learning material of TaRL; (ii) intervention (i) plus teacher training and mentoring by Pratham staff; (iii) intervention (i) and (ii) plus remedial activities by volunteer outside of school hours at the community (only for Bihar); and (iv) intervention (i), (ii) plus volunteer support for teachers within school hours (only for Uttarakhand) (Banerjee et al., 2016). In Bihar, they also planned a summer camp targeting children in the treatment group with lower proficiencies in basic reading before the beginning of the school year. The experiment in Bihar and Uttarakhand investigated which set of interventions was effective in changing the teaching practices of government school teachers.

It was found that the simple provisions of teaching and learning materials, and teacher training and mentoring did not change the behavior of teachers (Banerjee et al., 2016). Volunteers who supported teachers in regular classes did not organize remedial activities in basic reading and math; instead, they supported teachers in conducting regular classes. However, the teachers in Bihar organized the summer camp with TaRL, which demonstrated a significant improvement in learning in a period of less than four weeks (Banerjee et al., 2016). Teachers implemented TaRL successfully in the summer camp but not during the regular school day. This highlighted the question why they did not implement TaRL during the regular school day (Banerjee et al. 2017). Pratham and researchers referred to the interview records of teachers, and examined the reasons. Teachers found the material looked good but they did not have a time to utilize it since they had to finish the prescribed syllabus (Banerji, 2019). Pratham also noticed that while teachers were well trained in the intervention, other stakeholders in the government system, such as cluster coordinators, block resource persons, and district-level master trainers, were not involved in the activities of providing mentoring and support to teachers (Banerji, 2019).

In the experiment in Bihar and Uttarakhand, the research agenda for Pratham was "how it was possible to insert TaRL into a large poorly functioning government school system." From the findings in the experiment in the two states, Pratham refined the research agenda to "how to get government teachers to not only use materials and deliver the pedagogy, but also how to incorporate the targeted teaching aspect of the model into the regular school day" (Banerjee et al., 2017). Pratham developed an intervention strategy for government schools, based on the findings from Bihar and Uttarakhand. In Bihar, from 2012 to 2013, Pratham trained the local education officers called "cluster coordinators" on TaRL (Banerji, 2015). Cluster coordinators were originally in charge of administrative tasks for 12 to 15 schools but were not involved in academic guidance or support for teachers (Banerji, 2015). Through training and on-site practice, cluster coordinators became leaders of practice. They trained teachers and monitored the teaching practices at schools. An hour and a half were officially allocated within the school day for remedial activities. The pilot in the government schools succeeded in improving learning (Banerji, 2015). The model, composed of the involvement of supervisors as leaders of practice, and the allocation of time slot for TaRL in the school day, was called "the government partnership model." The model was later rigorously evaluated in an experiment in Haryana (Banerjee et al., 2017), which confirmed its positive impact on student learning.

6.5 Scaling-up TaRL by Pratham staff: experiment in Uttar Pradesh (2013-2014)

In the experiment in Bihar from 2008 to 2009, the summer camp demonstrated significant improvement in learning in a period of less than four weeks (Banerjee et al., 2016), which indicated that short bursts of remedial activity could improve learning. The experiment in Bihar targeted regular teachers, but the results provided insight for Pratham to review their own practices. It forged new research agenda of "how much time one Pratham staff member should spend in each school to bring about a tangible improvement in learning" (Banerji, 2019). Pratham developed a model in which one staff was in charge of two to three schools and rotated each school by six to ten days (Banerji, 2019). The model called "Learning camp model" was rigorously evaluated in an experiment in Uttar Pradesh (Banerji, 2019), which confirmed positive impact on student learning in basic reading and math.

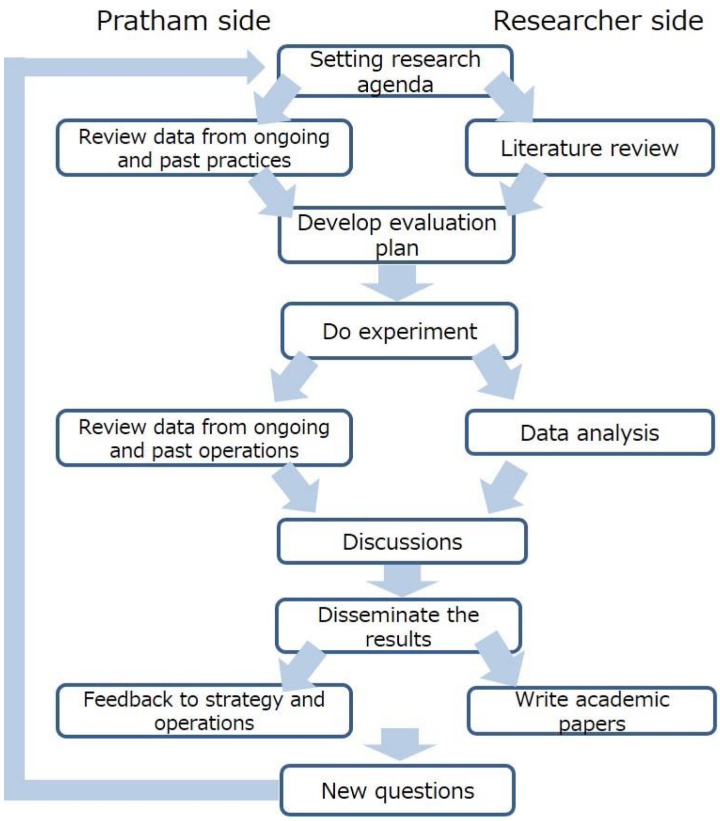
Through a series of experiments, Pratham and researchers developed two complementary strategies to scale up TaRL in rural areas: "the government partnership model" and "the learning camp model" (Banerjee et al., 2017). Pratham can demonstrate the impact of TaRL on foundational skills using the learning camp model. Referring to the results, Pratham discusses the plan of scaling up TaRL with the state government, and scales up TaRL by the government partnership model. The two models were mainstreamed in the Read India program (Pratham Education Foundation, 2017).

6-6. Process of the impact evaluation in the case of Pratham

Sections 6-3 to 6-5 in this chapter overviewed a series of experiments for a "search" of a better strategy for scaling up TaRL. Figure 1 depicts the common pattern in the process of the impact evaluations. The process began by setting the research agenda that addressed a gap between the mission of the organization, "Every child in school and learning well," and the actual performance level of Pratham. In the experiment of Jaunpur, Pratham set the research agenda "how to scale up TaRL so that it could benefit a large number of children in rural areas." In later experiments, Pratham refined the research agenda based on the findings in the past experiments. To enhance evidence-based policymaking and practice, Shar et al. (2015) argued that impact evaluations should be conducted on the demand from practitioners. The experiments in the case of Pratham were "demand-driven," as the research agenda was set from the strategic concern of the organization. Based on the research agendas, the evaluation plans were designed in detail.

The results of the experiments demonstrated the causal impacts of the interventions; however, the area of the experiments was geographically limited. Pratham triangulated the results of the experiment with the data and experience from their ongoing programs in other areas. Other aspects than the causal impact such as cost and time needed to be considered when developing a strategy to scale up TaRL in rural areas (Pratham, 2021).

Figure 6.1 Process of the impact evaluation in the case of Pratham



Through an experiment, Pratham and researchers learned whether and how the interventions worked in the field. They identified a new research agenda for the subsequent experiment. Rukmini Banerji, CEO of Pratham, stated that "[o]ur work together has been fueled not only by the need to know 'what works' but also guided by what is the new question that we need to explore to make the current work even stronger" (Pratham, 2021). Through several cycles of experiments, Pratham established two models for scaling up TaRL, which were integrated into their strategy and practices. A series of experiments was the process of "organizational learning" for Pratham.

The motivation for Pratham to conduct a series of experiments originated from the mission statement "Every Child in School and Learning Well." In 2005, at the time of the experiment in Jaunpur, Pratham felt frustrated by the gap between the mission and actual performance (Banerji et al., 2005). Pratham was not satisfied with the operation in the following four aspects, (i) the pace of learning gains in their program, (ii) efficiency in terms of time-use in class, (iii) knowledge sharing on pedagogy among staff, and (iv) insufficient impact on the government system. Pratham strived to fill the gap between the mission and actual level of performance; thus, the mission of Pratham was a driving force for the organization to conduct a series of experiments. The mission also anchored the consistency of the measured outcomes across several rounds of experiments. As the experiments consistently measured basic reading and math, Pratham and researchers could see progress in the model development.

6.7 Lessons learned for other development agencies

The case of Pratham shows three lessons regarding evidence-use by development agencies. First, setting the research agenda that addressed the gap between the mission and the actual level of performance was a critical initial step for the successful use of evidence. The experiment on the Balsakhi program in Mumbai confirmed the impact of TaRL on student learning. Considering the status of educational development in the country, Pratham set the research agenda "how to scale up TaRL so that it could benefit a large number of children in rural areas," then refined it in a series of experiment. Since the research agenda originated from the strategic concerns of Pratham, the results of the experiments were fed back well to strengthen the strategy and operations of the organization. The research agenda in the case of Pratham is similar to the concept of the "learning agenda," discussed in evidence-based policymaking and practice (United States Commission on Evidence-Based Policymaking, 2017). A learning agenda is "a set of broad questions directly related to the work an agency conducts that, when answered, enable the agency to work more effectively and efficiently, particularly pertaining to evaluation, evidence, and decision-making. (USAID, 2021)" The research agenda that Pratham set in a series of experiment was the critical learning agenda for the organization to achieve the mission.

Second, the leading management in Pratham played a central role in promoting and coordinating the experiment. To enhance evidence-use in international development, Shah et al. (2015) argued that impact evaluation should be embedded within the operational and decision-making structure of development agencies. Since the experiment in Mumbai, Rukmini Banerji, CEO from 2015, led the overall process of a series of experiments at the Pratham side. Setting a research agenda that addresses the gap between the mission and the actual performance of the organization required the understanding of the overall situation of programs and the contexts including the status of educational development and government policies. The progress of operations as well as the capacity of the Pratham staff varied across regions. Understanding the overall situation and context was also essential for selecting sites (state and/or district) that matched the research agenda. It was also necessary to triangulate the findings in experiments with data from the operations at other sites and other aspects such as cost and time for the better feedback into the strategy and operations. While the chief management position of operations in development agencies should take a leading role to utilize evidence to improve the programs, one of the typical problems for development agencies is staff turnover (Jones, 2012). To avoid loss of organizational memory due to frequent personnel changes, personnel arrangements such as cross-appointments should be taken for experienced and knowledgeable staff members to be continuously involved in research.

Third, the equal partnership between Pratham and researchers was the foundation to make the experiment productive. Banerji Rukmini was at first skeptical and suspicious for researchers, Abhijit Banerjee and Esther Duflo, when she met them in 1999 (Banerji, 2019); however, during the experiment on the Balsakhi program in Mumbai, they overcame several technical difficulties through mutual communication, and developed a trust between them (Banerji, 2019). Pratham had knowledge of the local contexts and expertise in the operations, and the researchers had knowledge of academic literature and expertise in scientific research. The experiment provided them with a shared context and an opportunity to exchange their knowledge and expertise. The protocol of the experiment also became a framework for their work and discussions. If the partnership was not equal, a successful exchange of knowledge and expertise between the two entities would not have occurred (International Initiative for Impact Evaluation, 2019). The three lessons from the case study above are fundamental, and provide insights for bilateral or multilateral development agencies to utilize impact evaluations, improve and expand their operations in education aid.

Reference

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics*, 122(3): 1235–1264.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster and Stuti Khemani. 2010. "Pitfalls of Participatory Program: Evidence from a Randomized Evaluation in Education in India." *American Economic Journal: Economic Policy*, 2:1: 1-30.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobini Mukerji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India." *NBER Working Paper,* No. 22746.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobini Mukerji, Marc Shotland, and Michael Walton. 2017. From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives*, Vol. 31, No. 4, Fall 2017: 73-102.

Banerji, Rukmini, Madhav Chavan, and Usha Rane. 2005. "Learning to Read." *Changing English*, 12:2: 177-190.

Banerji, Rukmini and Madhav Chavan. 2016. Improving literacy and math instruction at scale in India's primary schools: The case of Pratham's Read India Program." *Journal of Educational Change*, 17: 453-475.

Banerji, Rukmini. 2015. "How Do Systems Respond to Disruptive Pedagogic Innovations? The Case of Pratham in Bihar." *RISE Working Paper,* October 2015, RISE.

Banerji, Rukmini. 2019. "Banerjee and Duflo's journey with Pratham." *Ideas for India,* Nov. 13, 2019.

Banerji, Rukmini. 2020. "Can communities mobilize for schooling and learning?: Bottom-up perspectives from Pratham in India." In M. Nishimura (Ed.), Community participation with schools in developing countries: Towards equitable and inclusive basic education for all. New York & Oxon: Routledge.

Center for Global Development. 2006. *When will we ever learn? Improving lives through impact evaluation*. Center for Global Development.

Dhliwal, Iqbal, and Caitlin Tulloch. 2012. "From research to policy: using evidence from impact evaluations to inform development policy." *Journal of Development Effectiveness*, 4:4: 515-536.

Duflo, Esther. 2020. "Field Experiments and the Practice of Policy." *American Economic Review*, 2020, 110 (7): 1952-1973.

Dutt, Shushmita Chatterji, Christina Kwauk, and Jenny Perlman Robinson. 2016. *Pratham's Read India Program: Taking small steps toward learning at scale*. Center for Universal Education.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press.

International Initiative for Impact Evaluations. 2019. Using evidence to improve children's foundational skills: a successful teaching and learning approach expands in India and beyond. 3iE.

Jones, Harry. 2012. "Promoting evidence-based decision-making in development agencies." *Background Note*. Overseas Development Institute.

J-PAL 2017. Evaluation Summaries by J-PAL of Pratham's Education Programs 2001-2016. Retrieved from https://prathamusa.org/website-admin/wp-content/uploads/2017/04/JPAL.pdf.

Langer, Laurenz, Ruth Stewart, Yvonne Erasmus, and Thea de Wet. 2015. "Walking the last mile on the long road to evidence-informed development: building capacity to use research evidence." *Journal of Development Effectiveness*, 1095783.

Nutley, Sandra, Isabel Walter, and Huw Davies. 2007. *Using Evidence: How research can inform public services*. Bristol: Policy Press.

Pratham 2020. "Teaching at the Right Level: From concern with exclusion to challenges of implementation." *Background paper prepared for the 2020 Global Education Monitoring Report*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000373668/PDF/373668eng.pdf.multi

Pratham 2021. Replies to the questions from Maruyama. Unpublished.

Pratham Education Foundation 2017. *Annual reports 2016-17*. Pratham Education Foundation.

Pratham Education Foundation 2019. *Annual reports 2018-19*. Pratham Education Foundation.

Sabet, Shayda Mae and Annette N. Brown. 2018. "Is impact evaluation still on the rise? The new trends in 2010–2015." *Journal of Development Effectiveness*, 10:3: 291-304.

Shah, Neil Buddy, Paul Wang, Andrew Fraker, and Daniel Gastfriend. 2015. "Evaluations with Impact: Decision-focused impact evaluation as a practical policymaking tool." *3ie Working Paper* 25.

Simon, Herbert. 1979. "Rational Decision Making in Business Organizations." *American Economic Review*, Vol. 69, No.4: 494-513.

USAID 2021. Learning Lab: CLA Toolkit. Retrieved from https://usaidlearninglab.org/qrg/learning-agenda.

United States Commission on Evidence-Based Policymaking. 2017. *The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking*. Washington D.C., United States Commission on Evidence-Based Policymaking.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise.* Washington D.C.: World Bank.