INSTITUTO DE ECONOMÍA

DOCUMENTO de TRABAJO

487

2017

Tailoring Instruction to Improve Mathematics Skills in Preschools: A Randomized Evaluation

Francisco Gallego, Emma Näslund-Hadley, Mariana Alfonso

www.economia.puc.cl • ISSN (edición impresa) 0716-7334 • ISSN (edición electrónica) 0717-7593

TAILORING INSTRUCTION TO IMPROVE MATHEMATICS SKILLS IN PRESCHOOLS:

A RANDOMIZED EVALUATION*

Francisco A. Gallego* Emma Näslund-Hadley** Mariana Alfonso***

This Version: June 2017

Abstract

Previous research suggests that tailoring instruction to each student needs can produce significant learning gains. However, few programs have successfully implemented this approach in practice. In this paper, we present the results of a randomized evaluation of a program that uses an individualized scaffolding approach during regular school hours to teach the basic elements of numbers and shapes to preschoolers using a sample of 107 preschool centers and almost 3,000 children in Peru. The program improves Math outcomes among all children (by 0.10 standard deviations) and has stronger impacts for students in the lower quintiles of the distribution of outcomes and for students with teachers with university degrees. The effect in the areas that were implemented in a more intense way persists even one year after the program ended. Interestingly, we find no evidence of effects that are different across gender, language-spoken at home, and proxies for SES, contrasting with results from previous research that suggest that effects of Math programs are biased along gender and socioeconomic lines.

Keywords: Education, Mathematics, Early Childhood Development, RCT, Scaffolding.
JEL Classification: I21, I28, O15.

* Francisco A. Gallego, Pontificia Universidad Católica de Chile and J-PAL. email: fgallego@uc.cl

** Emma Näslund-Hadley, Inter-American Development Bank. e-mail: emman@iadb.org

*** Mariana Alfonso, Inter-American Development Bank. e-mail: marianaa@iadb.org

Mimate is a program designed and implemented by the Apoyo Institute in coordination with the Ministry of Education of Peru (MINEDU) and the Inter-American Development Bank (BID). We thank Inés Levy, Sergio de Marco, and Alejandro Saenz for excellent research assistance. We benefited from comments from Samuel Berlinski, Jessica Goldberg, Elizabeth Spelke, and audiences in seminars at Harvard, PUC-Chile, and the Workshop on "Equity: Contributions from research in Education, Economy, Psychology and Neurosciences" held at PUC-Chile in March 2016. Innovations for Poverty Action (IPA) assisted with the sampling, design field work, and statistical analysis of the data. We are also grateful to the Inter-American Development Bank, J-PAL, and Fondecyt (Grant # 11070181) for financial support. The usual disclaimer applies.

1. **Introduction**

The relevance of early childhood development and pre-school education for subsequent development of cognitive and non-cognitive abilities and economic and social performance is well known. (Cunha and Heckman, 2007). One of the key mechanisms to explain this effect is related to dynamic complementarities in which skills developed at early ages complement the future production of new skills. The previous argument is also true for the development of math skills, as mathematical thinking is cognitive foundational (Clements and Sarama, 2009, 2011). Poor mathematics performance throughout primary school can be traced to weaknesses in pre-primary number competencies (e.g., Gersten and Flojo, 2005; Malofeeva et al. 2004; National Mathematics Advisory Panel, 2008). Moreover, several studies document that math skills acquired in kindergarten predict academic performance in primary education better than early reading skills, verbal and spatial cognition, or even measures of attention span, memory, and social skills (e.g., Duncan et al., 2007, Romano et al., 2010, Jordan and Locuniak, 2009, Pianta et al., 2008a, Aunola et al., 2006). Still other studies find impacts of early math learning on adult life outcomes (e.g., Geary et al, 2013).

However, it is less clear what specific pre-school programs can improve learning for preschoolers and affect subsequent learning (Duncan and Magnuson, 2013; Berlinski and Schady, 2015; Clements and Sarama, 2011). Recent research suggests that changes in pedagogy –with the same stock of teachers—in primary schools can lead to significant improvements in learning levels, especially in the case of programs focused on targeting the teaching to each student needs—the “teaching at the right level” approach (e.g., Banerjee et al., 2016). One way to implement this idea is the *scaffolding* approach. Scaffolding is an adaptive interactive strategy that recognizes the current capacities of children (trainees) and guides them to further learning without creating frustration. Several papers document that some of the most promising recent models of parent-child, teacher-child, and parent-teacher-child relationships involve attachment and scaffolding as major determinants of child learning (Heckman and Mosso, 2014). Activities are tailored to the individual child’s ability so they are neither too hard nor too easy in order to keep the child in the “zone of proximal development”, which is the level of difficulty at which the child can learn the most (Heckman and Mosso, 2014). In the same way that a scaffold is used as a temporary structural support during building construction, scaffolded instruction provides targeted and temporary support to help students develop new skills and abilities (Englert et al., 1991).

In this paper we study the Mimate program, which uses an individualized scaffolding approach to teach the basic elements of numbers and shapes to preschoolers in Peru. The program seeks to change the traditional vertical teaching model of memorization and repetition into a horizontal, student-centered model in which students progress at their own rate.[1] The program includes sessions during the regular schedule of Peruvian preschools in which the teacher splits up the children into small groups or pairs for activities. The curriculum and lesson plan proceed with math challenges, gradually progressing from very basic to advanced, in which each task prepares the student to tackle the next one. Mimate emphasizes the development in two main areas: numerical literacy and understanding shapes. Each student also receives a personal box with hands-on teaching materials. All items are suggested to be kept in the "Mimate corner" of the classroom so that children can play with these toys and tools at free hours during the day. The program also includes twice-a-month formative assessments with a simple five minute round of flash cards between the teacher and individual students. Based on the students' answers, the teacher then knows which skills the students need to practice and can direct them to an appropriate activity. In addition, there is an initial training period for the teachers before the program starts and visits from teacher assistants during to ensure the quality of the implementation of the program. The cost of the program is about $150 per kid (in 2013 US dollars) in addition to the regular educational expenses.

The randomized controlled trial (RCT) was implemented in schools located in three cities of Peru: Huancavelica, Angares, and Ayacucho. We implemented a stratified randomization with 54 treatment schools and 53 controls schools considering the following strata: cities and urban/rural. The baseline information we collected suggest that the randomization was successful, as most variables were balanced across treatment and control groups.

We find that the intervention increased Math outcomes for children that attended the treated schools at the end of the program by 0.10 standard deviations (σ hereafter) for items related to numeracy and by 0.12σ for items related to understanding shapes (0.12σ). In addition, when one looks at specific items, we find that the impacts are bigger for: geometric shapes (with an impact of 0.20σ), number selection (with an impact of 0.18σ), additive composition (with an impact of 0.15σ), and oral counting (with an impact of 0.11σ). The effects in these dimensions are statistically significant even when correcting p-values for multiple hypotheses testing (using a family-wise error correction approach, as suggested by following

[1] Algan et al. (2013) document the presence and consequences for social capital of teaching practices that emphasize vertical teaching methods in contrast with group work and a more horizontal teaching practice in several countries of the world.

Anderson, 2008, which we use in all the results in the paper in which we test effects on different dimensions).

We also study the potential existence of heterogeneous effects along several dimensions related to characteristics of the students and the schools. First, we find that the impact of the program is stronger for children in the lower quantiles of the learning distribution using quantile regressions. This result contrasts with previous evidence that tended to find stronger effects for students with good math performance (e.g., see the reviews by Clements and Sarama, 2011 and Duncan and Magnuson, 2013). These results suggest that the scaffolding approach works in a better way for kids with lower abilities, who probably receive much more attention than in regular classrooms. Second, we study interaction effects considering several observable characteristics of the students and schools, such as: gender, language and socioeconomic status of the students and location of the school, class size, and teacher education. We do not find that the treatment has statistically heterogeneous effects in all these dimensions with the only exception of teacher education: we find that when the teacher who implemented the program has a university education, the average effect of the program increases to about 0.15σ (and the effect of the program for teachers without university education is not different from 0). This suggests that there is a complementarity between teacher human capital and a the program, which is probably related to the fact that a program like this may demand certain level of teacher human capital in order to be able to scaffold the learning process.

Next, we study the medium-term effects of the treatment using information on the same outcomes for the same kids one year *after the program ended*. Our results imply that the effects decrease in magnitude after one year. In particular, the treatment effects for the overall test and the test including numeracy are not different from 0 and treatment effects for the test that measures understanding shapes are equal to 0.06σ. This result resembles most of the previous results in the literature on pre-school interventions: significant, short-term effects on learning skills that decrease after the programs end (Duncan and Magnuson, 2013). The only item in which we find a big and statistically significant medium-run effects is on geometric shapes with an impact of 0.16σ, again correcting for multiple hypotheses testing. As we discuss below the bigger impacts on this area may be a consequence of the intensity of the program implementation in this area.

In terms of heterogeneous effects in the medium run, mirroring our short-term results, the only interaction effect we identify is related to a positive effect for the students that received the Mimate program with teachers with university education. Based on quantile regressions, we find that in the medium run the positive effects we estimate for understanding shapes seem to be concentrated in the students in the upper part of the distribution of outcomes. This result may suggest that when the students are left alone (after the program ended), those with more skills deal with new learning in a better way. More research is needed to have a clear answer to this point.

Finally, we also study whether the program has an impact on other dimensions of the learning process. On the one hand, there may be potential positive externalities of a program like Mimate if the new abilities of the students also affect child development and learning in other areas. On the other hand, there may be negative effects if the effort to improve Math outcomes displaces resources (such as time, teacher effort or even student interest) from other activities and learning dimensions. We do not find evidence of indirect effects using results of the Raven test of cognitive ability and using results of a writing test.

In order to help us with the interpretation of the results, we also collected information on the actual implementation of the program using administrative data, class observations, and surveys applied to teachers and parents. The results of a process evaluation imply that about 66% of the planned sessions were actually implemented. It is worth noting that this partial implementation was mainly due to the existence of a national teacher strike that cancelled up to three months of class time in some schools. This may explain the results we find because, in contrast to the sessions about shapes and figures that were covered at the beginning of the program, those associated to number sequence, quantities, and patterns came at the end of the program and were covered in a less intense way, accordingly to our process evaluation. Actually, our process evaluation implies that while 82% of the actual sessions planned to cover figures were implemented (equivalent to about 30 sessions implemented), just 57% of the sessions related to numbers were implemented (equivalent to about 25 sessions).[2] This may explain our results both in the short- and medium-run. In turn, results of the class observations and teacher responses to surveys suggest that the program affected the behavior students and teachers in the classroom, the beliefs of teachers about students, and the ability of teachers to meet the objectives and to teach Math (all of this with respect to the teachers in the control group). In contrast, teachers in the treated group do not report significant differences in terms of the availability of resources and materials covered in comparison to the control

[2] This is a very coarse classification between numbers-related and shapes-related sessions as the program seems to influence both dimensions in an interlinked way but we want to illustrate the differences between dimensions in sessions covered by the program.

group. This suggests that the intervention operated mainly through teaching practices more than through increases in available resources.

Although there are large literatures exploring the causal impacts of preschool programs (e.g. see the review in Almond and Currie, 2010) and exploring the effect of pedagogical innovations (e.g., see the review in Kremer et al., 2013) only a few studies implemented randomized controlled trials to study the effects of implementing pedagogical innovations among preschoolers in the developing world (for instance He et al., 2009 and Naslund-Hadley et al., 2014).[3] Thus, this paper makes several contributions to the existing literature. First, this study represents one of the few randomized evaluations of Math programs and the only randomized evaluation that combines a relatively big sample of schools and students and follows up the kids after the program ended in order to identify medium-run effects. Second, this paper also contributes to the research that emphasizes the adaptation of the teaching and learning process to each student. This is related to both the "teaching at the right level" (TaRL, hereafter) (e.g. Banerjee and Duflo, 2010; Banerjee et al., 2016) and the scaffolding (e.g., Heckman and Mosso, 2014) empirical literatures.[4] Regarding the TaRL literature, the Mimate program is closer to the implementations that focus on interventions based in the classroom with the current stock of teachers (e.g., He et al., 2008; Banerjee et al., 2010, Barrow et al., 2009, and Muralidharan et al., 2017). Regarding the scaffolding literature, while several papers have studied scaffolding programs in different contexts (ie., Lynch and Kim, 2017; Kim et al., 2017), to our knowledge, this study represents the first randomized evaluation of the scaffolding approach to Math instruction in developing countries. Third, we present evidence using a combination of process and impact evaluation to suggest that the part of the program that was implemented as planned – related to instruction about shapes – tends to have bigger effects in the short-run and some persistent effects in the medium-run (especially related to the dimension of geometric shapes). Moreover, the information collected from teacher surveys suggest that the program affected more strongly aspects related to pedagogy than to increases in resources or changes in the material covered in class. This is an important

[3] We do not discuss in detail the literature on the effects of early childhood interventions from both developed and developing countries that track individuals from early childhood into adulthood and show that children brought up in a more favorable early environment are healthier and taller, have higher cognitive ability and educational attainment, and earn significantly higher wages (see e.g., Paxson and Schady 2010; Stith, Gorman and Choudhury 2003; Liddell and Rae 2001; Walker, et al. 2005; Gertler, et al. 2014; Havnes and Mogstad 2011).

[4] Both approaches are related in the sense that they try to implement specific ways of teaching that adapt to each student level. However, a relevant difference among them is that in the scaffolding approach the specific way of adapting to each student is more focused in the relationship and interaction between the teacher and the student.

contribution to the literature, which typically does not present detailed information about the actual implementation of programs.

The remainder of the paper is structured as follows. Section 2 briefly presents the educational context of Peru and the preschool sector in that country and a description of the Mimate program and its relationship with the literature on math preschool programs. Section 3 describes the research design and the data used in the paper. Section 4 describes the implementation of the program using administrative data from the program, class observations, and answers to a teacher survey applied after the program was implemented. Section 5 discusses the main results of the paper including both short- and medium-term effects. Finally, Section 6 provides a summary of our findings and draws some policy implications.

2. Context and the Mimate program

a. Education in Peru

Access to education in Peru is high (net school enrollment in primary and secondary education is above 90% and 75%, respectively) but learning outcomes are not good and present steep socioeconomic gradients. At 15 years of age, according to the PISA report in 2012, Peruvian students scored the worst in mathematics, reading, and science out of all 65 countries tested. Similarly, according to the 2008 SERCE report, third-graders and sixth-graders in Peru also rank among the bottom five for the same subjects, and display alarming gaps between urban and rural, and male and female performance. In turn, using information from tests applied to the complete population of second graders of Peru, results imply a difference of about one standard deviation between children of the richest and poorest quintiles (Berlinski and Schady, 2015). Schady et al. (2015) document similar differences for six-year olds in urban and rural contexts of Peru. Even at five years of age, the participants in this study scored on average 15 percentage points lower on the Raven test of cognitive ability than five-year olds in the average of countries with data available (See Table 1). This difference is sizeable as it is equivalent to about 60% of a standard deviation observed for five-year olds in the UK.

The Peruvian Ministry of Education (MINEDU) has recognized pre-primary school as a priority to improve educational outcomes. From 2007 to 2011, MINEDU increased pre-primary education spending per student by 70%, compared to increases of 60% and 46% in primary and secondary education respectively (ESCALE, MINEDU). In parallel, the rate of attendance to preschool for children ages 3 to

5 years increased from 53% to 75% in the 2001-2012 period, with increases of attendance for children living in rural and urban areas (attendance increased from 53% to 75% in urban areas and from 44% to 66% in rural areas during the same period). Moreover, Belinski and Schady (2015) report that the gap in enrollment between richest and poorest quintiles dropped from 36% to 12% between 2000 and 2013.

In terms of pedagogical models in pre-school education, APOYO (2012) describes the status-quo as a situation in which there is a long tradition of memorization and repetition drills based on the same approach for all the students in the class and on a model of vertical education, in which all the students receive the same material and tend to repeat, at the same time, what the instructor does (e.g., all students sing at the same time or repeat the number the instructor mentions in front of the class).

b. The Mimate Program

Mimate was designed by pedagogical experts with the relevant literature on early education kept closely in mind including three 45 minute sessions per week fit into the daily schedule of Peruvian preschools. Mimate uses scaffolding techniques which include workbook, formative assessments, and visits from teacher assistants who ensure the quality of teacher-student interactions.

The aim of scaffolding IS to keep the student engaged in what Vygotsky (1978) described as the “zone of proximal development,” the range of concepts between what learners can do on their own and what can be achieved in collaboration with instructors or peers (Ellis and Worthington, 1994). Scaffolding can be achieved using soft or hard approaches (Saye and Brush, 2002). Soft scaffolding is human support provided by a teachers or peers that helps students meaningfully participate in the performance of actions. Teachers who provide soft scaffolding must remain cognizant of students’ stages of learning in order to provide just the right amount of support at the right time to each student (de Grave et al. 1999; Hogan and Pressley 1997; Lepper et al. 1997). Hard scaffolds are computer or paper-based materials that anticipate the student’s needs during the unit by following a pedagogical flow that builds upon concepts incrementally (Saye and Brush, 2002). Ideally, both soft and hard scaffolding measures can occur simultaneously to allow teachers to keep an entire classroom engaged with the material. Mimate uses both elements.

Individualized instruction requires teachers to know exactly where each child stands in terms of his or her understanding of the material.[5] This is no simple task in classrooms of up to 30 students per teacher and without access to technology. Mimate developed a solution that makes accurate, formative assessments possible with a simple five minute round of flash cards between the teacher and individual student.[6] Based on the student's answers, the teacher then knows which skills the student needs to practice and can direct them to an appropriate activity.

The nature of the program also needed teacher training, which included two different dimensions. Before the program started, teachers participating in the program received three training sessions in which they received information on the objectives of the program and training on the sessions and the materials to be used in the program. 83% of the teachers attended these sessions. The second component of training support corresponds to on-school training in which specialists from Instituto Apoyo observed the sessions taught by the teachers and gave advise on how to improve the application of the program. These sessions took place during the implementation of the program (from March to November). In average, teachers received six visits.[7]

In terms of Math contents, Mimate is motivated by research that suggests pre-schoolers can in fact learn Math and, moreover, most kids at the preschool age actually enjoy the basic math challenges (e.g., Seo and Ginsberg , 2004; Clements and Sarama, 2008). Similarly, research also shows that pre-school Math programs aimed at children from low SES can have significant impacts in their Math outcomes (e.g., Ramani and Siegler, 2011).[8]

The curricumlum of Mimate emphasizes the following four areas:

---

[5] This is related to literature that emphasizes the importance of the quality of the interactions between students and teachers in order to attain good learning outcomes (See Berlinski and Schady, 2015 and the references therein).

[6] Recently, sophisticated computer programs have become popular tools to test students and recommend further exercises based on their specific capabilities (e.g., see Barrow et al., 2009; Muralidharan et al., 2017); however, these programs were not feasible to implement in the context of Mimate where only seven percent of schools have internet access and computer skills are sparse.

[7] In addition, during the final part of the implementation of the program (between October and December), participating teachers were invited to attend inter-learning sessions in which they could share their experiences with other teachers. There were four of these sessions and in average 63% of the teachers attended these sessions.

[8] This is particularly relevant in the context of this study, as GDP per-capita in Peru was just about $11,000 (at PPP) in 2012 and the average years education of the adult population was about 8.5 (Barro and Lee, 2013). It is well known that low-income children enter kindergarten with fewer number-related experiences – at home, in their communities, and in their preschools – than their middle-income counterparts and that these early learning opportunities set middle-income children at an advantage (Clements and Sarama, 2008).

a) *Numerical literacy: number sequence and number aspects.* Children learn the variety of purposes for numbers: to count one-by-one, to order objects, to measure size or time, etc. The two basic types of "number words" (i.e. one, two, three etc. and first, second, third etc.) are also practiced.
b) *Numerical literacy: capturing structure.* Children practice the rule of cardinality, that the last number spoken when counting signifies the total number. When counting is understood, children are stretched to consider numbers in small groups and learn how to combine and separate groups mentally without relying on counting. This catalyzes quick and efficient mental calculus later on.
c) *Understanding shape: variety of geometric shapes.* Children play with various shapes (circles, squares, rectangles, triangles and rhombuses), different types of lines (straight or curved, closed, or open). They are challenged to visualize points, planes, and three dimensional spaces.
d) *Understanding shape: training fine motor skills.* Children construct shapes with clay, draw shapes with pencils, and learn symmetry by folding or cutting paper. Neuroscience has demonstrated that knowledge of form comes through experiential learning with the hands in one channel and language in the other. The activities hone the coordination of fingers while reinforcing the lessons of earlier modules.

3. **Research Design and Data**

A randomized controlled trial employing randomization at the school level forms the basis of the study design. We randomly selected 53 treatment schools and 54 control schools from Huancavelica, Angares, and Ayacucho. The control group would proceed with their existing plans (i.e., as we discussed above with teaching practices based on group instruction and on memorization and repetition) while the treatment group would adopt the Mimate program in their preschool classrooms. The randomization was implemented at the school level and considers the existence of six strata that come from the combination of the school location (urban/rural) and the city in which the school is located.

The timeline of the study is as follows. Baseline data were collected in March 2012, then we collected data at the end of the school year in December 2012 (that we call short-term follow up in the paper), and then we collected a new follow-up survey in December 2013 one year after the treatment was implemented and when the students were at the end of their first grade of primary education (that we call medium-term follow-up in the paper). In total, 2,400 children participated in the baseline and short-term tests, while 2,416 participated in the baseline and medium-term tests.

The primary data used in this paper were collected through the baseline and two follow-ups that we mentioned above. The general format of all these processes was very similar, consisting of application of tests in the baseline and two follow-ups and also the application of parent surveys and teacher surveys in the baseline and first follow up survey. The quantitative tests were preschool adapted versions of the "Early Grade Mathematical Assessment" (EGMA) originally developed by the Research Triangle Institute International. The test measures various abilities related to mathematics with the following exercises: Comparing quantity, basic shape recognition, advanced shape recognition, basic object counting, advanced object counting, number selection, advanced number selection, fine motor skills, symmetry, shape sequence, number sequence (clock), number sequence (calendar), additive composition, geometric shapes, and addition and subtraction word problems. Appendix 1 provides a detailed description of these items. In several exercises of the paper we present results of the items included in EGMA at three different levels of aggregation: (i) an overall index considering all the items included in EGMA, (ii) two indices including the items related to numerical and shapes abilities, (iii) measures of the development of each ability included in EGMA. In addition to tests related to measure the development of math skills, we also applied instruments to assess the development of non-math outcomes: the Raven test of cognitive development and writing tests.

The questionnaires applied to teachers and parents collected information about the child's classroom and home experience. We use the information in these surveys with two purposes. First, we use answers in the baseline survey to study heterogeneous effects of the treatment (e.g. size of classroom, access to materials, education level of teacher, education level of the mother, dominant language at home, etc.). Second, we use answers the teachers and parents surveys to understand the mechanisms through which the program may have affected learning outcomes.

In addition, to help the interpretation of the results, we also collected process information about the actual implementation of the program in all the treatment schools and we also observed 44 randomly selected schools (37 treatment and 7 control schools) in a qualitative way collecting measures of classroom infrastructure, teacher-student interactions, and student-student interactions between the months of October and December of 2012. Mathematics classes were videotaped and analyzed using the CLASS

(Class Assessment Scoring System), which codes behavior in three areas: emotional support (i.e. generating a respectful environment and listening to children), classroom organization (i.e. managing time and keeping control of students), and teaching support (i.e. developing concepts thoughtfully and reinforcing student learning) (See Pianta et al., 2008b for a description and Araujo et al., 2016 and Cruz-Aguayo, 2015 for applications in Latin American countries).

4. **Program Implementation**

This section presents information to characterize the actual implementation of the Mimate program in several dimensions. This is important because of three reasons. First, it allows us to know the actual fidelity of implementation of the treatment. Second, while comparing with the teaching of regular schools (ie., those in the control group), it allows us to understand the intensity of the different components that may be affected by the program. Third, this also allows to understand in a better way the status-quo of teaching in the control group. This information can gives us useful insights to interpret the quantitative results of the program. We use information from three different sources that we now discuss separately.

a. **Process Information**

In parallel to the experimental evaluation, we collected data on the actual implementation of the program. The objective of this process was to produce rough measures of the actual implementation of the program both in terms of quantity and quality.

First, we collected administrative data on the actual implementation of all the sessions by school. Table 2 presents a general description of how the program was planned from March to December. The total program considered 86 sessions. The administrative data implies that, in average, 66% of the sessions were implemented. During 2012, there was a teacher strike that cancelled up to three months of class time in some schools. In total, 87% of teachers in the sample participated in the strike at some point in time. This rate was higher (95%) in rural areas than urban areas (83%).

The partial implementation of the program implied that the actual coverage of the sessions planned to be taught in the last three months of the year was much smaller than the topics covered at the beginning of the year. For instance, while all the sessions planned from March to July were implemented by all the schools, none of the topics planned for November and December were covered. Given the intended order of the sessions, this partial implementation implied that the actual coverage of the numerical literacy and the understanding shapes parts of the program were significantly different. In fact, while 82% of the sessions considering materials directly related to shapes were implemented, just 57% of the sessions directly related to numeracy were covered. This is an important feature that we will consider in our empirical analyses.

In order to study whether there is some systematic pattern of implementation related to observable characteristics of the schools, in Table 3 we present a regression analysis in which we study the correlation of the percentage of sessions covered in each treatment school and a number of observable characteristics of the students (i.e., baseline math scores, gender, language, and socioeconomic status) and schools (location, average class size, and teacher education). In columns (1) to (7) we present the relation of the percentage of completed sessions with each variable individually and in column (8) we include all the variables together. Our models are not that successful in terms of predicting the number of sessions actually implemented. All the variables combined can just explain about 22.5% of the variance of the number of sessions (column 8). In terms of individual significance, we find that schools with higher test scores at the baseline, with more Spanish speaking students, with more students coming from high SES households, and located in urban areas have more sessions. However, in all the cases, the size of the impact, while statistically significant, is not economically relevant. For instance, the estimated coefficients imply that a school with students with math scores one standard deviation above the average school implemented just 5 percentage points more sessions than schools with average students. Similarly, rural schools decrease the percentage of implemented sessions by just 5.7 percentage points with respect to urban schools and schools with only bilingual students reduce the share of implemented sessions by just 8.1 percentage points with respect to schools with just Spanish speaking students. Moreover, when including all the variables together (column 8), none of them has a correlation that is statistically different from 0 (probably a consequence of the high collinearity among the right-hand side variables). In all, these results suggest that while some variables seem to be correlated with the percentage of completed sessions, none of these variables seem to have a first order effect on the actual implementation of the program.

b. Class observations

Next, in order to study the "quality" of the implementation of the program we use class observations from a sample of 44 schools (37 from the treatment group and 7 from the control group). The idea was to identify a number of relevant dimensions and see whether the actual implementation was consistent with the objectives of the program.

We now report several results that come from these class observations. First, in terms of the "Mimate corner," all the schools in the treatment group have the Mimate corner in the classroom and in 78% of the cases the actual design of the corner was consistent with the design suggested by the program. Second, in terms of the actual use of the materials, in 100% of the cases, teachers were using the materials as suggested by the program. Third, in terms of allowing the students to explore and manipulate the materials used in the learning process, 94% of the teachers in the Mimate program allowed kids to manipulate the materials but there are some differences among urban (100% allowed kids to manipulate the materials) and rural schools (78% of the teachers allowed the kids to manipulate the materials). Fourth, as suggested by the design of the program, in terms of allowing the kids to discover by themselves the objectives of the activities, in most cases teachers were always or often doing that (treated schools have an average of 3.5 in a metrics that goes from 0 to 4 with the following categories: 1=never, 2=sometimes, 3=often, and 4=always). Fifth, the board games provided by Mimate were available to the students in 92% of the schools. Finally, in 100% of the treated schools, boys and girls were included in activities equally. All these results suggest that the fidelity of the implementation of the program in the sessions developed was very high.

In terms of comparison with the control schools using the qualitative class observations we identify several interesting facts. First, the Mimate classrooms achieved 38% higher marks on the assessment of whether the class was "prepared and structured with a clear objective". Teachers in the control group on the other hand were many times observed improvising with activities during the lesson. Second, teachers implementing the Mimate program seemed to have more patience with their students. Mimate teachers were more consistently observed encouraging kids to try activities multiple times in a friendly manner (in 95% of the cases), compared to the control group (with 71% of cases). Furthermore, Mimate teachers were observed to be "paying attention to those students who do not understand well and explaining with patience" at a higher rate (95%) than the control group (63%). This implies that the Mimate program seemed to have helped head teachers organize their classrooms, use the time in a more efficient way trying to reach to students that lagged behind, and adopt a more patient attitude.

In addition, using CLASS we identify the performance of different classes in terms of the following three dimensions (using scales that go from 1 to 7): (i): the measure of emotional support implied a higher average for treatment schools (5.32 for treatment schools 4.78 for control schools, which is equivalent to about 0.72 standard deviations of the variable), (ii) the measure of classroom organization implied also a small difference in favor of the treatment schools (4.47 for treatment schools versus 4.30 for control schools, equivalent to 0.16 standard deviations of the variable), and (iii) the measure of teaching support also presented a small difference in favor of the treatment schools (3.93 for treatment schools versus 3.79 for control schools, equivalent to about 0.15 standard deviations of the variable).[9] Thus, in terms of these comparisons, we do see some differences "in favor" of treatment schools (with the caveat that the small sample size makes very difficult to make statistical inference from these observations).

c. Teachers and parents follow-up surveys

We use answers to surveys applied to 101 teachers in 91 schools and 1,780 parents in 99 schools after the treatment was applied. We use this data as a source of additional information on the actual implementation of the program. Given that we also include control schools in this sample, we are able not only to understand the changes created by the program but also the counterfactual situation without the treatment.[10]

In terms of data coming from the teachers survey we find several significant differences between the treatment and control groups. In terms of the teaching of Math, teachers in the treatment are more likely to agree that they have enough time to teach all the material they were supposed to cover (71% in the treatment group versus 48% in the control group). At the same time, they are more likely to disagree with the statement that children get tired when learning Math (32% in the treatment group versus 11% in the control group). This may be a consequence of the fact that they are more likely to strongly agree that they teach using games that include all students in the class (55% among treated teachers versus 38% among control teachers). In terms of preparation, treated teachers are more to likely to strongly agree that they are accurately prepared to teach Math (23% in the treatment group versus 7% in the control group). They are also more likely to strongly disagree with liking the idea of moving to teach in another educational

[9] As a reference, Cruz-Aguayo et al (2015) report results of the application of the CLASS instrument in classrooms in grades K-2 in Brazil, Chile, and Ecuador. Their results also imply a similar order in terms of results to what we observe in our study: worst performance for the area of teaching support and similar results for emotional support and classroom organization.

[10] All the differences between the treatment and control group we report are statistically different from zero (using standard errors clustered at the school level).

level (42% in the treatment group versus 17% in the control group). Interestingly, at the same time teachers do not report significant changes in terms of infrastructure and resources to teach Math and in terms of changing objectives in the terms of the materials covered in class. This suggests that the Mimate program helped teachers to manage in a better way the classroom and the contents more than to an increase in pedagogical resources per-se.

Teachers also report significant changes about their students. Teachers in treated schools are more like to report that : (i) most of their students behave well in class (68% in the treatment group versus 50% in the control group), (ii) most of their students have the capacity to learn both Math (46% in the treatment versus 29% in the control group) and also any material they are taught (55% in the treatment group versus 39% in the control group), (iii) more than half of their students do not have problems at home that affect their learning (81% in the treatment group versus 53% in the control group), (iv) just a minority of students do not have support from their parents (15% in the treatment group versus 30% in the control group), (v) just a minority of students have concentration problems while in class (17% in the treatment group versus 32% in the control group), and (vi) most of their students have good performance in Math (55% in the treatment versus 36% in the control group). Thus, teachers improve their opinions about the abilities, performance, and the social context of their students.

All Math programs are subject to challenge on how they solve the potential gender-bias present in teaching the subject. Interestingly, teachers in the Mimate program have more equalitarian beliefs about happens in the classroom and about their students. The answers to the survey imply that teachers in treatment schools are more likely to report that there are no differences among children of both genders in terms of: (i) discipline problems (35% in the treatment group versus 18% in the control group), (ii) demands for attention of the teacher (67% in the treatment group versus 45% in the control group), (iii) paying attention in class (49% in the treatment group versus 38% in the control group), (iv) following instructions in class (67% in the treatment group versus 39% in the control group), and (v) performance in Math (80% in the treatment group versus 54% in the control group). These results, especially the last one, suggest that the program was able to produce a more gender-unbiased approach to the believes of teachers and to the teaching and to the learning of Math. This is important to understand the results in terms of learning we present below.

Finally, we also use the parents surveys after treatment to understand whether the program changed the beliefs, attitudes, homework, and involvement of parents in their children education. We do not observe significant differences among parents in treatment and control schools in all these dimensions. The only exception is that parents are more likely to strongly agree that their kids will do well on Math in school (41% in the treatment group versus 33% in the control group). This suggest that the program does not seem to have affected parenting in a significant way and, therefore, it is unlike that this dimension is relevant to explain the impact estimates we present below.

In sum, the process information reported in this section suggests that even when the program had a partial implementation related to teacher strikes, results from both class observations and teacher surveys do suggest that the program affected several margins related to the behavior of teachers and students in the classroom in a way consistent with the approach of the program.

5. **Empirical Results**

We start the discussion of the empirical results of this paper by characterizing the situation of the population included in the study at baseline. As discussed before, the baseline tests completed by all participants revealed a real academic disadvantage for preschoolers compared to international standards (see Table 1). In practice, this implies that their basic mathematics competence was concerning. For instance, fewer than four in ten children could write their age and only about half of the children could count to the number ten. As for the teachers, 12% had earned a Masters or Doctorate, 44% percent possessed a university degree, and the rest had earned a teaching certification without attending university. On average, teachers had 14 years of experience and were responsible for 24 students.

Table 4 presents an analysis of heterogeneity in mathematics outcomes across different groups using the baseline data. We present standardized outcomes (i.e. with an average of zero and a standard deviation of one) for tests that measure mathematics, cognitive ability, writing, and oral comprehension. We identify significant gaps in learning among most dimensions: urban vs. rural students; Spanish, Quechua, and bilingual speakers; and males vs. females. This is not surprising when one considers that, in terms of general conditions of the schools, less than half of the participating rural schools were connected to the public water system, compared to 88% of urban schools, and one in five had a computer, compared to one in two for urban schools. In this setting, rural students scored on average 0.46 standard deviations below their urban peers in the baseline assessment of mathematics.

Two in three of the participating rural students spoke exclusively Quechua in their home, and the rest were split evenly between bilingual or Spanish speakers. Only 2% of participating urban students spoke exclusively Quechua at home, and 99% spoke Spanish only. The Quechua speakers performed even lower than the rural group (0.61 standard deviations below Spanish-speaking students), and the bilingual students also suffered significantly (0.46 standard deviations below Spanish speaking students).

In the reminder of this section we present first the econometric models and statistical tests used in the paper, second an initial check of internal validity and attrition in the three data collection processes and, then, we move to present the main results of the experiment for both the short-term and medium-term follow-ups considering both average and heterogeneous treatment effects.

a. *Econometric Models and Statistical Tests*

The empirical framework for the main analyses of this paper uses observations on children who attend schools that were randomly assigned to the treatment and control groups. In particular, we estimate the following regression model:

$$Y_{ij} = \theta + \alpha T_j + X'_{ij}\beta + \varepsilon_{ij}, \tag{1}$$

where Y represents an outcome for student i who attends school j, T is a dummy variable that takes a value of 1 if the school is part of the treatment group, X is a vector of control variables, and ε is the error term that is clustered at the school level j in each regression. Then, estimates of α quantify the differences in means between students attending schools that receive the Mimate program over the school year 2012 and those that did not receive it.

We include the following variables, in vector X: the value of Y for student i at the baseline test, strata dummies, and dummies for test givers. If the randomization is successful, adding these variables should not change the estimate of α and these controls should just provide more precise estimates. For instance, it is well-known that tests scores at the student level present a huge-level of variation and, therefore, controlling for baseline estimates reduces the idiosyncratic variation of the results and, therefore, allows us to obtain more precise estimates (Duflo et al., 2008). We actually checked that the differences in estimates of α when estimating equation (1) with and without including X come basically from differences in the precision of the estimates of α. Therefore, in order to save space we only report estimates including X in the tables of the paper.

In addition we also implement the following additional econometric exercises to study the existence of heterogeneous treatment effect. First, we run regressions considering interaction effects of the treatment dummy and students/schools observable characteristics of the following form:

$$Y_{ij} = \theta + \alpha T_j + \gamma T_j * W_{ij} + \eta W_{ij} + X'_{ij}\beta + \varepsilon_{ij}, \qquad (2)$$

where W corresponds to a variable that varies at the school or student level. Then, if we estimate that γ is statistically different from 0, it implies that treatment effects are different for students having different W.

Second, we estimate quantile regressions, in which we estimate the effects of treatment for students in different quantiles of the distribution of outcomes, following a structure similar to the one presented in equation (1). In this case we obtain an estimator of α that is different for individuals in different quantiles of the distribution. In this way, we can estimate the effect of the treatment on the distribution of outcomes (conditional on outcomes in the baseline, as we include them in the vector of covariates).

We include in our study a number of different outcomes and therefore we may face an inference problem related to multiple hypotheses testing. Basically the problem is that significant coefficients may emerge simply by chance, even if there are no treatment effects (Anderson, 2008). Thus, we may be over rejecting the null-hypothesis of no treatment effect. Thus, in the case of tests about the existence of treatment effects on individual components of the test, we calculate adjusted p-values following the methodology suggested by Anderson (2008). In particular, this procedure corrects the p-value for the family wise error rate (the probability of making a Type I error) and uses a Westfall and Young (1993) type correction. We call these FWER p-values.

b. Balance and Attrition

Panel A of Table 5 compares outcomes at the baseline for students in the treatment and control groups. We do not find differences in six out of the seven variables we test. The only statistical difference we find is that the share of women among treatment group students (52%) is higher than in the control group (43%). Other variables do not present statistically significant differences. Among the variables without statistical differences, while the magnitude of the differences was very small in some cases (e.g., the shares students who attend preschool at ages 3 and 4, of rural school and Quechua speaking households), in other cases, the differences were bigger (in scores on baseline tests of Math, cognitive development, and writing). It is worth noting that in all the regressions we present in the paper we control for baseline test scores and we also present regressions in which we show that controlling for a gender dummy variable does not change the results in a meaningful way.

In terms of attrition, the main sample includes 2,926 students attending the schools included in the sample considered in this experiment. We were able to collect information for 2,400 students in the first follow-up in December 2012. This implies an attrition rate of 17.9 percent. The main reasons for not being found was that the student had withdrawn from the school or were absent when the instruments were applied. The attrition rate is not statistically different among treatment and control group students. Moreover, Panel B of Table 5 presents a re-analysis of balance of the sample for which we have information for the December 2012 follow-up. Results imply a very similar picture to what we saw in Panel A of Table 5 for the baseline: only the gender variable is unbalanced and the differences in the other variables present the same order of magnitude to what we saw in the baseline. This implies, that baseline characteristics of the attriting and non-attriting students are not significantly different by treatment assignment. Next, in terms of the second follow-up application of instruments in December 2013, we were able to collect information for 2,446 students, which implies an attrition rate of 16.4 percent. Panel C of Table 5 presents an analysis of balance for this survey and, as before, we find a similar pattern to what we observe in Panels A and B of Table 5.

c. Short-term Results

We start discussing results of the follow-up collected at the end of the 2012 school year (December, 2012). This corresponds to the end of the implementation of the Mimate program for the cohort of students for which we apply the baseline in March 2012.

First, we present impact estimates of the Mimate program on the overall EGMA test scores considering all the dimensions included. Table 6 presents the results in column 1. The estimate is statistically significant and implies an impact of the Mimate program of 0.10σ in favor of students in the treatment group. As a benchmark, Araujo et al. (2016) document that well identified "teacher effects" (i.e., the effect of increasing the quality of the teacher by one standard deviation) for a sample of kindergarten classrooms in Ecuador is 0.09σ. Thus, the impact effects are of the same order of magnitude of a huge improvement in teacher quality.

Next, we study the potentially different effects of the program in the numeracy and shapes abilities that the Mimate program aims to develop. This is important as our implementation analysis implies a much bigger progress in terms of the implementation of the sessions focused on the teaching of shapes than in the case of the sessions focusing in the teaching of numeracy skills. We present the results in columns 2 and 3 of Table 6. Results imply a slightly bigger impact (0.12σ) on an index that considers just the abilities related to shapes in EGMA than on an index considering just the abilities related to numeracy in EGMA (an impact of 0.10σ). This is consistent with the fact that the sessions related to shapes had a higher degree of implementation than the sessions related to numbers. At the same time, the teaching of numeracy was nearer from the measurement of skills than the teaching of shapes, which may imply stronger effects for numeracy.

In order to formally test for the effect of the share of sessions on learning outcomes, we run regressions of the effect of the overall EGMA score on the share of Mimate sessions actually implemented in each school (this is equal to 0 for all control schools). This is presented in column (4) in Table 5. We find that the coefficient of the share of sessions implemented is equal to 0.15 (significant at the 1% level). Obviously, these results assume that the actual number of sessions is not endogenous to other unobserved school characteristics. To test the last point, we present in column (5) of Table 6 IV estimates where we use both the treatment indicator and the interaction of the treatment indicator and the duration of the teacher strike in each school as instruments for the share of completed sessions.[11] Results imply an estimate of 0.14, which is statistically indistinguishable from the OLS estimate.[12] These results imply that an increase of the number of sessions from what we observed in average (about 57 sessions) to the planned number of sessions (86) would increase the effect of the program by 0.05. These results also imply that differences in the share of sessions related to numeracy and shapes (25%, accordingly to our process data) would imply a difference of 0.035 standard deviations in terms of the effects in both dimensions, which is above what we observe in estimates of columns (2) nd (3) of Table 6 (0.02 standard deviations). Thus,

[11] In the first stage we find that the coefficient of the treatment status is 1.01 (with a stadard error of 0.05) and for the interaction is -0.0041 (with a stadard error of 0.006). This implies that 20 additional days of the teacher strike decrease the share of sessions by 8.2%. Notice also that the average duration of the strike was 49 days for treatment schools.

[12] We also run an additional exercise in which control for all the school variables included in Table 3 as correlates of the share of completed sessions. Results in both the first- and second-stages do not change in a significant way (eg., the second stage estimate is now 0.138).

we interpret these results as suggesting that probably the number of completed sessions may explain the slightly bigger effect of the program on skills related to shapes.

Next, we present estimates for the items considered in the EGMA test. We do this in order to learn the dimensions of the math learning that are affected the most by the Mimate program. We present both naïve (without considering the potential existence of a multiple hypothesis testing problem) and corrected (considering multiple hypothesis testing) p-values in Table 7. First, in terms of numeracy skills, we find statistically significant effects for number selection (with an effect of 0.18σ significant at the 1% level), additive composition (with an effect of 0.15σ significant at the 5% level), and oral counting (with an effect of 0.11σ significant at the 5% level). Second, in terms of abilities related to shapes, we find a significant effect for geometric shapes (with an effect of 0.20σ significant at the 1% level).

We now present estimates to test the existence of heterogonous treatment effects in Table 8.[13] First, we study whether Mimate has a differential effect by gender. This is important because of several reasons: (i) there are gender gaps in learning outcomes at baseline (see Table 5), (ii) gender was not balanced across treatment and controls students in the baseline, (iii) several papers document big differences in terms of mathematics teaching across genders, with a bias towards boys, and (iv) the results of the teacher survey suggest that the program affected in a significant way the beliefs of teachers towards a more gender unbiased approach. We do not observe statistical differences of treatment effects between boys and girls. This implies, on the negative side, that Mimate was unable to close the existing gender gap but that, on the positive side, it is a program that is gender neutral, which is something important for the teaching of math and is consistent with the results we obtain with the teachers survey.[14]. Mimate inherently addresses this gender problem by requiring regular attention to each student, regardless of sex, and thus preventing the monopoly of attention that fidgety and high-energy boys can sometimes demand from their teachers at the expense of better-behaved girls. In fact, as we discussed before, (i) the qualitative findings indicate that in all schools using Mimate, boys and girls were included in activities equally, (ii) results in Table 3 suggest that the number of sessions implemented was not different in classes with more girls than in classes more boys, and (iii) the answers to the teacher survey strongly suggest a more gender-neutral approach to teaching among treated teachers.

[13] We present heterogeneous effects for the overall EGMA test in order to save space and because we do not see systematic differences in the pattern of heterogeneous effects when analyzing effects for the numeracy and shapes dimensions of the test.

[14] The existence of boys-biased programs in schools is not unfamiliar to Peru, results from a science program implied that the program widened the achievement gap between boys and girls (Beuermann et al., 2013).

Second, we study whether there are heterogeneous effects by the language of the student. This is a very challenging dimension in terms of the existing gaps at the baseline and also given the difficulty of adapting the materials to several languages. Results in column 2 of Table 8 imply that we do not observe statistical differences for students who speak Quechua or are bilingual. Actually, if anything, the treatment effects seem to be smaller for Spanish-speaking students. Thus, as in the case of gender, Mimate seems to be a program that is neutral in terms of outcomes. This is very remarkable as our implementation analysis implied that schools with more Spanish-speaking students tended to have more sessions of the program implemented.

Next, we study whether results are different for students from high and low SES households (column 3 of Table 8). As previously discussed, the exposure to different levels of mathematical reasoning for students from different SES families seems to have strong effects in terms of math learning. Yet the Mimate program did not just favor the privileged and well-educated sectors. Results imply that there are no statistical differences and that, if anything, children of low socioeconomic status improved at a faster rate than children from high socioeconomic status household. The same result emerges in terms of comparisons between students from rural and urban areas. These results echo the findings of Ramani and Siegler (2011), providing another example that fun, interactive mathematical games are also beneficial for low socioeconomic status students.

We also study whether the effects of the Mimate program depend upon some characteristics of the school. First, we consider potential differences between schools located in urban and rural areas (column 4 in Table 8). Results imply that we do not find statistical differences among students attending urban and rural schools and that, if anything, the effects for students attending rural schools are bigger than for students attending urban schools. Again, this is a striking result given the challenges that rural schools face and that we document before. We then consider the effect of class size (column 5 in Table 8). Teachers face the difficulty of keeping a large classroom focused on learning, but by emphasizing work in small groups, Mimate appeared to neutralize the effect of class size as well. The students from large classrooms improved equally, on average, compared to students from smaller classrooms. Finally, we study whether the human capital of the teacher implied different program effects. Results in column 6 of Table 8 imply that this is the case. Teachers with university titles have stronger effects than the other teachers. In fact, the effect of the Mimate treatment is 0.15σ for schools in which the treatment was implement by teachers with university titles in contrast to an effect of just 0.04σ (and not statistically different from 0) for schools in which the program was implemented by a teacher without a university degree. The inherent classroom difficulties in Peru along with a completely new curriculum and pedagogical model can explain why teachers who possessed teaching degrees saw their students improve significantly more with Mimate in the short-term than less educated teachers.[15]

In all, the results in Table 8 suggest that the Mimate program did not have much heterogeneous effects for students with different characteristics, which is good news in comparison to the status quo of big gaps in several dimensions, such as gender, language, income, and location. The only difference we found was related to the fact that Mimate programs seem to need teachers with high human capital to maximize its impacts on kids.

Next, we present the results of treatment effects estimated using quantile regressions. This exercise allows us to estimate the impact of the treatment on the distribution of test scores (Table 9). In mathematics especially, teachers often find it especially difficult to find a balanced lesson plan that stimulates the curiosity of the high-achieving students without confusing the low-achieving ones.[16] Mimate's scaffolding model, which adapts to meet the individual needs of students may help to mitigate this problem naturally, as also suggested by class observations and answers to the teacher survey. Results in Table 9 imply that the treatment effects are significantly stronger for students in the bottom part of the scores distribution than for students in the top part of the distribution. For instance, while the impact for students in the bottom 10% and 25% of scores improved by between 0.12σ and 0.16σ (depending on the test considered), treatment effects for students in the top 10 and 25% of the score distribution were just between 0.04σ and 0.08σ. This shows an equitable impact of the program and suggests that it may even narrow the knowledge gap between the most motivated and the most distracted students.

Finally, Table 10 presents treatment effects of the Mimate program on non-math dimensions of the learning process. We focus on the Raven cognitive development test and on a writing test. This way, we study whether there may be positive or negative externalities of the program. Results imply that the program did not have effects that are statistically different from 0 on those dimensions. This suggests that neither the additional learning of math skills created positive externalities on other areas nor the new focus on math skills and teaching did not distract resources from learning in other dimensions.

[15] It is interesting to notice that we also find that years of teacher experience did not have a significant impact on the effects of the program, suggesting that teachers with old habits were not slower to change and adopt the innovative Mimate model.

[16] There is a general literature on the effects of several innovations to try to deal with class heterogeneity, such as tracking (see Duflo et al., 2011 and the references therein) and the use of software that allows kids to advance at different speeds (see Barrow et al., 2009 and the references therein).

In sum, the short-term results imply that the Mimate program had a statistically significant effect on general learning outcomes with slightly stronger effects in terms of the development of abilities related to understanding shapes. In terms of heterogeneous effects, the program seems to have been very successful in terms of providing a program that did not produce differences against girls and students from low socioeconomic status families (in terms of language, income, and urban location). Also, we do not identify different effects for classrooms with different sizes but we do find stronger effects when teachers with university titles implemented the program.

d. Medium-term Results

We now move to the discussion of the results of the follow-up implemented at the end of the 2013 school year in December 2013. This corresponds to a follow-up after the student spent one year of regular classes *without* receiving the program. This way we study the medium-term effects of the intervention one year after the program ended. We present results for the same exercises that we implemented in the case of the short-term follow-up.

In terms of the effects of Mimate on the overall EGMA test, we do find effects that are not statistically different from 0 (Table 6, Panel B). This implies that the effect of the program that we found in the short-term does not translate to medium-term significant effects. This is not a new result in the literature on the effects of math pre-school programs in primary education (e.g. see Duncan and Magnuson, 2013). However, as we discussed above, given the big difference in terms of the implementation of the program in the areas related to shapes and numeracy and the consequent differences we find in their EGMA score counterparts, we also expect to have a differential effect of the Mimate program in both areas. Results in Panel B of Table 6 confirm this idea. We find that Mimate has an effect that is not statistically different from 0 for the numeracy score and an impact equal to 0.06σ (with a p-value of 0.053). In addition we also find a significant effect for geometric shapes with an effect of 0.15σ (and significant at the 5% level using the FWER p-values). This result is interesting as we find that the strongest learning area in the short-term is also the one that has the most effects in the medium-run and with a big effect. It is important to stress that we are very conservative in reaching this conclusion as we are correcting our p-values by multiple hypotheses testing and, therefore, it is highly unlikely that this result is just an outcome obtained by chance.

In terms of heterogeneous effects of the intervention, we do not find many significant effects, with the exception again of a positive and statistically significant interaction of the Mimate treatment for teachers with university titles. However, the total effect of the Mimate program is still not different from 0 for students receiving the Mimate program. In any case, it is still interesting that this interaction is statistically significant and has almost the same effect in both follow-ups (0.11σ in the short-term follow-up and 0.12σ in the medium-term) suggesting a potential complementarity between Mimate and the human capital of the teachers.

In the case of students located at different points in the distribution of outcomes (Panel B of Table 9), we find that the significant effects that we find for shapes are concentrated among the students located in the top part of the distribution of outcomes. This result suggests the existence of a specific type of complementarity in which when more able students are left alone (after the program ended) they benefit from the program in the medium-run more than low-ability students while in the short-run (when the program is working) the opposite is true. More research is needed to have a clear answer to this point.

6. **Discussion and Conclusions**

There is no silver bullet curriculum that solves all education problems, but tested pedagogical strategies backed by theory are the best hope for delivering the returns on investment in education (Clements and Sarama, 2011; Duncan and Magnuson, 2013). The importance of numeracy skills at the preschool level has become increasingly apparent with the emerging number of randomized controlled trials using vast samples in North America and Europe. In Peru, where underachievement in mathematics persists from preschool all the way to secondary school and the Ministry of Education is looking to break from the old routine of memorization and repetition, a well-designed program could replicate or exceed North American and European effect sizes. In theory, Mimate provides children with age-appropriate materials that can bolster their foundations in mathematics and trains teachers to become listeners and facilitators of each child’s development.

The results in this paper suggest that Mimate has statistically significant impacts after one-year of implementation and some persistent effects one year later. The effects seem to be slightly stronger in outcomes related to forms and shapes than to numeracy in the short-term and the effects are only persistent in the medium term for items related to shapes. This is consistent with the fact that the process information we present in this paper suggests that the implementation of the program was higher exactly in sessions related to this area. We do not find evidence that the implementation was significantly different for schools and teachers of different characteristics, suggesting that the variation in the implementation was closely related to the interruption of the program associated to teacher strikes. Thus, our results suggest that the heterogeneous implementation of the program is probably driving the different effects in the numeracy and the shapes outcomes. It remains to be seen whether a complete implementation of the program can improve the impacts of the numeracy items both in the short-run and in the medium-run.

In terms of cost-effectiveness, the complete program has an annual average cost of about \$150 per student per year (on top of the regular expenditures). Considering that just about 2/3 of the program was implemented and the impacts we estimate, this implies that \$100 increases about 0.1 standard deviations. If we compare this to some of the results reported by Kremer et al. (2013) for innovation for primary education in developing countries, we find that the program is less cost-effective than the most efficient innovations these authors review but it is more cost-effective than most of the interventions involving teachers that are reviewed in the paper.

There are several implications and questions for future research and experimentation that can be derived from our results. Dramatic changes, like asking experienced teachers to step down from the pulpit and provide tailored instruction based on each student's progress, can take some time to take hold. The results that better qualified, but not more experienced, teachers taught the program more effectively suggest that teaching flexibility, or tools of teaching learned later in school, are critical to the program's success. In this dimension, one could think of teacher training visits or direct special attention to less educated teachers as interventions to help close this gap. One could assume that further teacher training efforts would improve the teaching results of less educated teachers, but further research would be needed to establish this connection.

Another dimension that may be relevant is extending Mimate into the home with simple take-home games (e.g. dice, puzzles). This could increase the intensity of the treatment, especially in the case of teacher strikes. Parents would then have something tangible to demonstrate approval and encouragement with math. Role models, guardians, and parents can be key agents for the success of Mimate because, in the end, they are the most influential teachers of all and children who see the value of practicing and playing with mathematics will find it easier throughout their education.

Another possibility for extension would be to introduce the Mimate program to first graders, or even second graders because, in Peru, these children are struggling with the same concepts. Over half of second-graders could not perform first grade level tasks (e.g. basic addition or measuring with equal-sized units) (New York State P-12 Standards, 2012). Furthermore, less than seven percent of second-graders in the regions of Ayacucho, Huancavelica, and Angares performed up to the second-grade level, (e.g. understanding place value) (ESCALE, 2012).

References:

Algan, Y, P. Cahuc, and A. Shleifer (2013). "Teaching Practices and Social Capital," American Economic Journal: Applied Economics, American Economic Association, vol. 5(3), pages 189-210, July

Almond, Douglas, and Janet Currie. (2010) "Human Capital Development Before Age Five," Handbook of Labor Economics, 4B, 1316–1476.

Anderson, M. (2008) ""Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". Journal of the American Statistical Association 103(484): pp. 1481-1495.

APOYO (2012) "Propuesta de implementación de un programa educativo para fomentar capacidades matemáticas en niños y niñas de 5 años en aulas de Ayacucho y Huancavelica" Mimeo.

Araujo, C., P. Carneiro, Y. Cruz, and N. Schady (2016) "Teacher Quality and Learning Outcomes in Kindergarten", Quarterly Journal of Economics, 131(3).

Aunola, et. al. (2006) "Developmental dynamics between mathematical performance, task motivation, and teachers' goals during the transition to primary school". British Journal of Educational Psychology, 76, 21–40.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukherji, M. Shotland, and M. Walton (2016) "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India" NBER Working Paper No. 22746.

Banerjee, A. and E. Duflo (2010) *Poor Economics*. Public Affairs.

Barro, R. and J-W. Lee (2013) Barro-Lee Educational Attainment Dataset http://barrolee.com

Barrow, L., L. Markman, and C. Rouse, (2009), "Technology's Edge: The Educational Benefits of Computer-Aided Instruction," The American Economic Journal: Economic Policy, (1), 52-74.

Berlinski, S. and N. Schady (2015) *The Early Years Child Well-Being and the Role of Public Policy*. Interamerican Development Bank.

Beuermann, D., E. Naslund-Hadley, I. Ruprah, and J. Thompson (2013). "The Pedagogy of Science and Environment: Experimental Evidence from Peru," Journal of Development Studies, vol. 49(5), pages 719-736,

Clements D.H., Sarama J. (2008) Experimental evaluation of the effects of a research-based preschool mathematics curriculum. American Education Research Journal; 45(2):443–494.

Clements, D. H., Samara, J. (2009). Early Childhood Mathematics Education Research: Learning trajectories for young children. New York, NY: Routledge.

Clements, D. H., Samara, J. (2011). "Early Childhood Mathematics Intervention." Science 333(6045): 968 –70.

Cruz-Aguayo, Y., J. LoCasale-Crouch, S. Schodt, T. Guanziroli, M. Kraft-Sayre, C. Melo, S. Hasbrouck, B. Hamre, and R. Pianta. 2015. "Early Schooling Classroom Experiences in Latin America: Focusing on What Matters for Children's Learning and Development." Mimeo, IADB.

Cunha, F. and J. Heckman (2007) "The Technology of Skill Formation." American Economic Review, 97(2): 31-47.

de Grave, W. S., Dolmans, D. H. J. M., & van der Vleuten, C. P. M. (1999). Profiles of effective tutors in problem-based learning: Scaffolding student learning. Medical Education, 33, 901-906.

Duflo, E., R. Glennerster, M. Kremer (2008). "Using Randomization in Development Economics Research: A Toolkit" In T. Schultz and John Strauss, Eds., Handbook of Development Economics. Vol. 4.

Duflo, E., P. Dupas, and M. Kremer (2011). "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." American Economic Review, 101(5): 1739-74.

Duncan, G. et. al. "School Readiness and Later Achievement". Developmental Psychology 2007, Vol. 43, No. 6, 1428–1446

Duncan, G. and K. Magnuson. (2013). "Investing in Preschool Programs." Journal of Economic Perspectives, 27(2): 109-32.

Ellis, E.S., Worthington, L. (2004). Research Synthesis on Effective Teaching Principles and the Design of Quality Tools for Educators Technical Report No. 5 produced for the National Center to Improve the Tools of Educators, University of Oregon

Englert, C. S., Raphael, T. E., Anderson, L. M., Anthony, H. M., & Stevens, D. D. (1991). Making strategies and self-talk visible: Writing instruction in regular and special education classrooms. American Educational Research Journal, 2, 337-372.

ESCALE. Estadistica de calidad educativa. Ministerio de educacion, Peru. http://escale.minedu.gob.pe/

Geary , D.C., M. Hoard, L. Nugent, D. Bailey (2013). "Adolescents' Functional Numeracy Is Predicted by Their School Entry Number System Knowledge". PLOS ONE 8 (1).

Gersten R, Jordan N.C., Flojo J.R. (2005). Early identification and interventions for students with mathematics difficulties. Journal of Learning Disabilities;38(4):293–304. [PubMed: 16122059]

Gertler, P et al. (2014) "Labor Market Returns to Early Childhood Stimulation: A 20-year Follow-up to an Experimental Intervention in Jamaica", Science, 2014, Volume 344, Issue 6187.

Havnes, T. and M. Mogstad (2011). "No Child Left Behind: Subsidized Child Care and Children's Long-Run Outcomes," American Economic Journal: Economic Policy, vol. 3(2), pages 97-129, May.

He, F., L. Linden, and M. MacLeod (2009) "A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trail." Mimeo, University of Texas, Austin.

Heckman, J. and S. Mosso (2014). "The Economics of Human Development and Social Mobility," Annual Review of Economics, Annual Reviews, vol. 6(1), pages 689-733.

Hogan, K., & Pressley, M. (1997). Scaffolding scientific competencies within classroom communities of inquiry. In K. Hogan & M. Pressley (Eds.), Scaffolding student learning: Instructional approaches & issues (pp. 74-107). Cambridge, MA: Brookline.

Jordan, N. Huttenlocher, J., and S. Levine. (1992). Differential calculation abilities in young children from middle and low-income families. Developmental Psychology;28:644–653.

Jordan, N. Locuniak, M. (2009)Early Math Matters: Kindergarten Number Competence and Later Mathematics Outcomes. *Dev Psychol*. May; 45(3): 850–867.

Kim JS, Burkhauser MA, Quinn DM, Guryan J, Kingston HC, and Aleman K. (2017) "Effectiveness of structured teacher adaptations to an evidence-based summer literacy program." *Reading Research Quarterly*. Forthcoming.

Kremer, M., C. Brannen, and R. Glennerster (2013) "The Challenge of Education and Learning in the Developing World" Science, Vol. 340 no. 6130 pp. 297-300, April.

Lepper, M. R., Drake, M. F., & O'Donnell-Johnson, T. (1997). Scaffolding techniques of expert human tutors. In K. Hogan & M. Pressley (Eds.), Scaffolding student learning: Instructional approaches & issues (pp. 108-144). Cambridge, MA: Brookline.

Liddell, C. and G. Rae (2001) "Predicting early grade retention: A longitudinal investigation of primary school progress in a sample of rural South African children" British Journal of Educational Psychology 71 (3): 413–428,

Lynch K, Kim JS (2017). Effects of a summer mathematics intervention for low-income children. *Educational Evaluation and Policy Analysis*. Forthcoming.

Malofeeva E, Day J, Saco X, Young L, Ciancio D. (2004). Construction and evaluation of a number sense test with Head Start children. Journal of Educational Psychology; 96(4):648–659.

Muralidharan, M., A Singh, and AJ Ganimian (2017) "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India" NBER Working Paper No. 22923

Naslund-Hadley, E., S. Parker, J. Hernández (2014) "Fostering Early Math Comprehension: Experimental Evidence from Paraguay". Global Education Review 1 (4):

National Mathematics Advisory Panel. Foundations for success: The final report of the National Mathematics Advisory Panel. Washington, DC: U.S. Department of Education; 2008.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R. M., & Morrison, F. J. (2008a). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45(2),* 365–397.

Pianta, R. C., K. M. La Paro, and B. K. Hamre. (2008b). *Classroom Assessment Scoring System (CLASS) Manual: Pre-K*. Baltimore, MD: Brookes Publishing Co.

Paxson, C. and N. Schady, 2010. "Does Money Matter? The Effects of Cash Transfers on Child Development in Rural Ecuador," Economic Development and Cultural Change, vol. 59(1), pages 187-229, October.

Ramani, G. B. & Siegler, R. S. (2011). Reducing the gap in numerical knowledge between low- and middle-income preschoolers. Journal of Applied Developmental Psychology, 32, 146-159.

Romano E., Babchishin L., Pagani L.S., Kohen D. (2010). School readiness and later achievement: replication and extension using a nationwide Canadian survey. Dev Psychol Sep;46(5): 995-1007.

Saye, J. W., & Brush, T. (2002). Scaffolding critical reasoning about history and social issues in multimedia supported learning environments. Educational Technology Research and Development, 50(3), 77-96.

Schady, N., J. Behrman, M. Araujo, R. Azuero, R. Bernal, D. Bravo, F. Lopez-Boo, K. Macours, D. Marshall, C. Paxson and R. Vakis (2015) "Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries." Journal of Human Resources 50 (2).

Seo, K.-H., & Ginsburg, H. P. (2004). What is developmentally appropriate in early childhood mathematics education? Lessons from new research. In D. H. Clements, J. Sarama & A.-M. DiBiase (Eds.), Engaging young children in mathematics: Standards for early childhood mathematics education (pp. 91-104). Hillsdale, NJ: Erlbaum

Stith, A., Gorman, K.S & Choudhury, N. (2003). Beyond primary school: Psychosocial risk and gender. Applied Psychology: An International Review. 52(4), 614-629.

Vygotsky, Lev. (1978) Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Walker, S. P., S. M. Chang, C. A. Powell, and S. M. Grantham-McGregor (2005). "Effects of early childhood psychosocial stimulation and nutritional supplementation on cognition and education in growth-stunted Jamaican children: prospective cohort study." The Lancet 366 (9499), 1804–1807.

Westfall, P. H. & Young, S. S. (1993). Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. John Wiley and Sons

Appendix 1: Exercises included in the "Early grade Mathematical assessment" (EGMA)

The quantitative tests were preschool adapted versions of the "Early Grade Mathematical Assessment" (EGMA) originally developed by the Research Triangle Institute International. The test measures various abilities related to mathematics with the following exercises:

a) *Comparing quantity:* Children are tested for their understanding of "more," "less," and "equal" with an image exercise comparing rows of kittens, chickens, and bunnies. The tester challenges the child to indicate in the each row which box had more, less, or equal numbers of animals.

b) Basic shape recognition: Children are presented with a circle, triangle, square, and rectangle and asked to correctly name each shape.

c) Advanced shape recognition: The same shapes are dispersed in a matrix of shapes, and the children are asked to match four plastic tiles (a circle, square, triangle, and rectangle) with their corresponding shapes.

d) *Basic object counting:* Children are asked to point with their finger and count balloons in a picture of balloons numbered 1, 2, 3, 4.

e) *Advanced object counting:* The same exercise is repeated with a picture of twelve balloons numbered 1 through 12.

f) *Number selection:* Children view a grid of twelve boxes, each one containing a different number of stars rising from 1 to 12. Then, they are asked to point out which box has three stars, six stars, nine stars, and twelve stars.

g) *Advanced number selection:* Children view three boxes containing clusters of four hearts, five hearts, and eleven hearts respectively. Below the boxes are the numbers 5, 11, and 4 placed out of order. The children are asked to match each symbolic number with the appropriate box of hearts.

h) *Fine motor skills:* Children copy to the best of their ability images of basic shapes, symbolic numbers and letters. Scores were later calculated by a team of digitation specialists.

i) *Symmetry:* Children are shown an image of a butterfly and asked to draw a line over the butterfly that divides it into two equal parts.

j) *Advanced symmetry:* Children try to match one side of a house with one of three options to complete the picture.

k) *Shape sequence:* Unfinished patterns of shapes (triangle, square, circle, triangle, square…etc.) are presented to the children. Children indicate which shapes on the right side of the page would complete the pattern for each row. Patterns with color were also tested and color blindness did not affect performance.

l) *Number sequence (clock):* Confronted with an image of a clock face, children respond to the questions "Which number comes after 4?" and "Which number comes before 9?"

m) *Number sequence (calendar):* The tester shows children a calendar page of April and asks the children to help reschedule a party by answering the question "What day is two days after April 5th?"

n) *Additive composition:* On the left side of the page is a box displaying three kittens and on the right side are pairs of similar boxes showing another numbers of kittens. Children are asked to identify the correct pair of boxes (one kitten and two kittens) that together are the same as the box on the left. The exercise is repeated three more times with flowers, apples, and hearts.

o) *Geometric shapes:* Children are given four plastic triangle tiles and a plastic rhombus tile and asked to arrange them to cover up a large hexagon shape depicted on the page. Then, the tester takes away two triangles and gives the child a rhombus and asks them to complete the task again.

p) *Addition and subtraction word problems*: "Daniel has one dog. María has one dog. How many dogs do they have in total?" and "There are four children walking to school. Two of them are boys and the rest are girls. How many girls are walking to school?"

Table 1
Global Raven cognitive test scores

Country	Score	Standard deviation	Average Age	Source
Peru	12.3	5.5	5.2	This study
United Kingdom	15.1	3.7	5.3	Gathercole, 1994
Denmark	18.5	2.4	5	Olsen 1992
India	12	4.5	5	Bhogle 1992
Poland	15.1	4.5	5	Szustrowa 1992
Slovakia	14.4	3.8	5.7	Ferjencik 1992
Argentina	14	3	5	Leibovich de Figueroa 1992
Canada	14.2	-	5.3	Wright 1995
Global Average	14.5	3.8	5.1	

Notes: Raven test comparison between countries.

Table 2

Process Information: Planned and Implemented Sessions

Module (by date)	Number of sessions	Average of sessions completed (%)
	March	
Basic shapes	2	100.00%
Spatial relationships	7	100.00%
	April	
Bend and cut	4	100.00%
Cover figures I	3	100.00%
	May	
Construct with cubes	3	100.00%
Number sequence I	3	100.00%
Dice images	4	100.00%
	June	
Define quantities I	6	100.00%
Assemble quantities	2	100.00%
Order numbers I	4	100.00%
	July	
Quantity representations	3	100.00%
	August	
Draw figures and patterns	3	99.39%
Symmetry	5	95.27%
	September	
Cover figures II	2	83.64%
Pre-writing I	2	69.09%
Recognize shapes and space	4	50.91%
Pre-writing II	1	43.64%
Cognitive game	2	39.09%
	October	
Pre-writing III	1	32.73%
Bend	5	17.45%
Number sequence II	2	14.55%
Define quantities II	2	6.36%
Pre-writing IV	1	5.45%
	November	
Order numbers II	3	0.00%
Pre-writing V	1	0.00%
Number patterns	7	0.00%
	December	
Numbers for measuring	2	0.00%
Number representations	2	0.00%
Total number of sessions	86	
Total avg. % of completed sessions	65.96%	

Notes: Comparison between planned and actually completed number of sessions by module. The order of the modules in this table follows the MIMATE timeline. Data comes from the Mimate program.

Table 3
Determinants of the Percentage of Completed Sessions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline overall Math score	0.087**							-0.032
	(0.042)							(0.056)
% women per (treated) school		0.062						0.141
		(0.077)						(0.086)
% of quechua speaking households per (treated) school			-0.067**					-0.031
			(0.029)					(0.045)
% of bilingual speaking households per (treated) school			-0.081**					-0.083
			(0.036)					(0.051)
% of high income families per (treated) school				0.067*				0.006
				(0.034)				(0.050)
Urban schools					0.057**			0.044
					(0.022)			(0.040)
Average class size per school						0.003*		-0.000
						(0.002)		(0.003)
% of teachers with university degree per (treated) school							0.020	0.038
							(0.026)	(0.027)
Observations	43	43	43	43	43	43	40	40
R-squared	0.121	0.012	0.142	0.070	0.122	0.096	0.014	0.225

*Notes: Dependent variable is the percentage of completed sessions per (treated) school, which is an average of the percentage of completed sessions of all the (treated) classes in each school. Each column shows the correlation between the dependent variable and some potential determinants. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1*

Table 4
Educational outcomes by groups (standardized effects)

	Boys	Girls	Rural	Urban	Spanish	Quechua	Bilingual
Mathematics	0.02	-0.04	-0.37	0.09	0.15	-0.46	-0.31
Cognitive ability	0.04	-0.05	-0.28	0.08	0.15	-0.36	-0.28
Writing	0.01	-0.01	-0.59	0.16	0.22	-0.64	-0.42
Oral comprehension	0	-0.01	-0.4	0.11	0.17	-0.44	-0.36
N	1529	1397	623	2303	1809	393	243

Notes: Baseline average deviation from the sample mean of each subgroup for different outcomes. Standardized effects for the complete sample. Data comes from the Mimate program.

Table 5

Sample Balance

	Treatment	Control	Difference	t-stats
Panel A: Baseline survey				
Score on baseline math	-0.04	0.03	-0.07	-0.83
Score on baseline cognitive	-0.08	0.08	-0.16	-1.31
Score on baseline writing	-0.07	0.08	-0.15	-1.24
Female	0.52	0.43	0.09	3.00
Attended preschool age 3 and 4	0.86	0.84	0.02	0.66
Speaks Quechua at home	0.19	0.18	0.01	0.08
Attends a rural school	0.21	0.21	0.00	-0.04
Panel B: Short-term follow-up				
Score on baseline math	-0.04	0.01	-0.05	-0.49
Score on baseline cognitive	-0.10	0.01	-0.11	-0.88
Score on baseline writing	-0.08	0.02	-0.10	-0.75
Female	0.51	0.44	0.07	2.33
Attended preschool age 3 and 4	0.87	0.82	0.05	1.21
Speaks Quechua at home	0.19	0.20	-0.01	-0.21
Attends a rural school	0.21	0.24	-0.03	-0.38
Panel C: Mid-term follow-up				
Score on baseline math	-0.03	0.04	-0.07	-0.71
Score on baseline cognitive	-0.07	0.07	-0.14	-1.14
Score on baseline writing	-0.05	0.09	-0.14	-1.14
Female	0.51	0.44	0.07	2.46
Attended preschool age 3 and 4	0.87	0.83	0.04	1.04
Speaks Quechua at home	0.19	0.19	0.00	0.00
Attends a rural school	0.21	0.21	0.00	-0.08

Notes: Data comes from the Mimate program.

Table 6
Effects on Mathematics Outcomes

Panel A: Short-term effect

Dependent Variable	Overall Test	Shapes Items	Numeracy Items	Overall Test	Overall Test
	(1)	(2)	(3)	(4)	(5)
Treatment	0.10***	0.12***	0.10***		
	(0.03)	(0.04)	(0.03)		
Share of Completed Sessions				0.15***	0.14***
				(0.05)	(0.05)
Controls	Yes	Yes	Yes	Yes	Yes
Estimation Method	OLS	OLS	OLS	OLS	IV
Observations	2400	2400	2400	2152	2152

Panel B: Medium-term effect

	Overall Test	Shapes Items	Numeracy Items	Overall Test	Overall Test
	(1)	(2)	(3)		
Treatment	0.00	0.06**	-0.03		
	(0.03)	(0.03)	(0.05)		
Share of Completed Sessions				0.00	0.00
				(0.02)	(0.02)
Controls	Yes	Yes	Yes	Yes	Yes
Estimation Method	OLS	OLS	OLS	OLS	IV
Observations	2416	2416	2416	1895	1895

*Notes: Standard errors corrected for heteroskedasticity and intra-cluster correlation at school level in parenthesis * p < 0.1, ** p < 0.05, *** p < 0.01. Data comes from the Mimate program.*

Table 7

Effects by items of the EGMA test

Panel A: Short-term effect				Panel B: Medium-term effect			
Dependent variable	Coefs.	Naïve p-value	FWER p-value	Dependent variable	Coefs.	Naïve p-value	FWER p-value
Understanding shapes							
Geometric shapes	0.195	0.000	0.000	Geometric shapes	0.159	0.002	0.030
	(0.0371)				(0.0522)		
Shape recognition	0.143	0.016	0.127	Shape recognition	0.0691	0.267	0.516
	(0.0592)				(0.0623)		
Spatial ability	0.130	0.022	0.147	Spatial ability	-0.0773	0.192	0.516
	(0.0566)				(0.0592)		
Symmetry	0.0985	0.113	0.397	Symmetry	0.0933	0.059	0.309
	(0.0621)				(0.0494)		
Figure sequence	0.0688	0.257	0.527				
	(0.0607)						
Reproduce figures	0.0669	0.384	0.527	Reproduce complex figures	0.0663	0.185	0.516
	(0.0768)				(0.0500)		
Numerical Literacy							
Number selection	0.178	0.000	0.011	Number selection	0.00914	0.893	1.000
	(0.0489)				(0.0678)		
Additive composition	0.148	0.002	0.036	Additive decomposition	-0.0246	0.694	0.999
	(0.0476)				(0.0625)		
Oral counting	0.105	0.002	0.036	Oral counting	-0.018	0.644	0.999
	(0.0340)				(0.039)		
Addition and subtraction word problems	0.118	0.012	0.112	Addition and subtraction problems	0.000527	0.992	1.000
	(0.0469)				(0.0560)		
Comparing quantity	0.0962	0.066	0.350	Comparing quantity	-0.0849	0.187	0.865
	(0.0523)				(0.0643)		
Naming numbers	0.0939	0.085	0.358	Naming numbers	-0.0724	0.267	0.916
	(0.0545)				(0.0652)		
Advanced numeration	0.0818	0.135	0.432	Advanced numeration	0.0151	0.749	0.999
	(0.0547)				(0.0473)		
Number sequence	0.0704	0.159	0.435	Ordering numbers	-0.0501	0.442	0.986
	(0.0500)				(0.0651)		
Comparing numbers	0.0525	0.224	0.448	Comparing numbers	-0.0613	0.315	0.945
	(0.0431)				(0.0610)		
Basic numeration	0.0202	0.660	0.681	Number sequence	0.0298	0.662	0.999
	(0.0461)				(0.0682)		
				Previous and subsequent	-0.00556	0.934	1.000
					(0.0672)		
				Measurement units	-0.0701	0.216	0.880
					(0.0567)		
				Writing numbers	-0.0323	0.651	0.999

Notes: Naive p-values correspond to standard p-values of the individual regressions. FWER p-values correct for the familywise error rate, (the probability of making a type I error) and use a Westfall-Young (1993) type correction as explained by Anderson (2008). Data comes from Mimate program.

Table 8
Heterogeneous effects: Overall index

Subgroup(s):	Female (1)	1-Quechua 2-Both (2)	High socio-economic level (3)	Urban (4)	Section size (5)	Teacher w/ university title (6)
Panel A: Short-term effect						
	Mathematics					
Treatment	0.11***	0.08*	0.11***	0.15**	0.13	0.04
	(0.04)	(0.04)	(0.04)	(0.07)	(0.10)	(0.05)
Subgroup 1 x Treatment	-0.00	0.02	-0.06	-0.06	-0.00	0.11*
	(0.03)	(0.09)	(0.06)	(0.08)	(0.00)	(0.06)
Subgroup 2 x Treatment		0.07				
		(0.07)				
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2400	2037	1149	2400	2400	1901
Panel B: Medium-term effect						
	Mathematics					
Treatment	-0.01	0.00	0.00	0.01	0.01	-0.05
	(0.04)	(0.04)	(0.04)	(0.08)	(0.10)	(0.05)
Subgroup 1 x Treatment	0.00	-0.01	0.04	-0.00	-0.00	0.12*
	(0.04)	(0.10)	(0.05)	(0.09)	(0.00)	(0.06)
Subgroup 2 x Treatment		0.01				
		(0.10)				
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2416	2202	1091	2416	2293	1811

Notes: All columns control for baseline score, school locality, test giver. Only column (2) considers two subgroups and two interactions: Quechua speaking household and household that speak Quechua and Spanish equally. Section size in column (5) is a continuous variable. Standard errors corrected for heteroskedasticity and intra-cluster correlation at school level in parenthesis. $* p < 0.1$, $** p < 0.05$, $*** p < 0.01$. *Data comes from the Mimate program.*

Table 9
Effects according to baseline mathematics quantiles

Panel A: Short-term effect

	Overall	*Shapes*	*Numbers*
	(1)	(2)	(3)
q10 Treatment	0.12***	0.15***	0.16***
	(0.04)	(0.05)	(0.05)
q25 Treatment	0.14***	0.15***	0.13***
	(0.03)	(0.03)	(0.04)
q50 Treatment	0.09***	0.13***	0.09***
	(0.02)	(0.02)	(0.02)
q75 Treatment	0.08***	0.08***	0.05**
	(0.03)	(0.02)	(0.02)
q90 Treatment	0.04	0.05**	0.05**
	(0.02)	(0.02)	(0.02)
Controls	Yes	Yes	Yes
Observations	2400	2400	2400

Panel B: Medium-term effect

	Overall	*Shapes*	*Numbers*
	(1)	(2)	(3)
q10 Treatment	-0.04	0.02	-0.04
	(0.04)	(0.03)	(0.04)
q25 Treatment	-0.01	0.02	-0.03
	(0.03)	(0.03)	(0.03)
q50 Treatment	-0.02	0.03	-0.01
	(0.03)	(0.02)	(0.02)
q75 Treatment	0.03	0.07***	0.00
	(0.02)	(0.03)	(0.02)
q90 Treatment	0.04	0.10***	-0.01
	(0.02)	(0.11)	(0.02)
Controls	Yes	Yes	Yes
Observations	2416	2416	2416

*Notes: Standard errors corrected for heteroskedasticity and intra-cluster correlation at school level in parenthesis. * p < 0.1, ** p < 0.05, *** p < 0.01. Data comes from the Mimate program.*

Table 10

Effects on Other Learning Outcomes

	Cognitive abilities	Writing
	(1)	(2)
Panel A: Short-term effect		
Treatment	0.06	-0.02
	(0.08)	(0.05)
Controls	Yes	Yes
Observations	2387	2365
Panel B: Medium-term effect		
Treatment	0.030	0.031
	(0.059)	(0.060)
Controls	Yes	Yes
Observations	2,111	2,068

*Notes: Standard errors corrected for heteroskedasticity and intra-cluster correlation at school level in parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Data comes from the Mimate program.*