IDB WORKING PAPER SERIES Nº IDB-WP-1388

Improving Early Literacy through Teacher Professional Development:

Experimental Evidence from Colombia

Horacio Alvarez Marinelli
Samuel Berlinski
Matías Busso
Julián Martínez Correa

Inter-American Development Bank
Department of Research and Chief Economist

October 2022

Improving Early Literacy through Teacher Professional Development:

Experimental Evidence from Colombia

Horacio Alvarez Marinelli*
Samuel Berlinski**
Matías Busso**
Julián Martínez Correa**

* World Bank
** Inter-American Development Bank

October 2022

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Improving early literacy through teacher professional development: experimental evidence from Colombia / Horacio Alvarez Marinelli, Samuel Berlinski, Matias Busso, Julian Martinez Correa.
p. cm. — (IDB Working Paper Series ; 1388)
Includes bibliographic references.
1. Early childhood education-Colombia. 2. Effective teaching-Colombia. 3. Interaction analysis in education-Colombia. I. Alvarez, Horacio. II. Berlinski, Samuel, 1970- III. Busso, Matías. IV. Martinez-Correa, Julián. V. Inter-American Development Bank. Department of Research and Chief Economist. VI. Series.
IDB-WP-1388

http://www.iadb.org

Copyright © 2022 Inter-American Development Bank. This work is licensed under a Creative Commons IGO 3.0 Attribution-NonCommercial-NoDerivatives (CC-IGO BY-NC-ND 3.0 IGO) license (http://creativecommons.org/licenses/by-nc-nd/3.0/igo/legalcode) and may be reproduced with attribution to the IDB and for any non-commercial purpose, as provided below. No derivative work is allowed.

Any dispute related to the use of the works of the IDB that cannot be settled amicably shall be submitted to arbitration pursuant to the UNCITRAL rules. The use of the IDB's name for any purpose other than for attribution, and the use of IDB's logo shall be subject to a separate written license agreement between the IDB and the user and is not authorized as part of this CC-IGO license.

Following a peer review process, and with previous written consent by the Inter-American Development Bank (IDB), a revised version of this work may also be reproduced in any academic journal, including those indexed by the American Economic Association's EconLit, provided that the IDB is credited and that the author(s) receive no income from the publication. Therefore, the restriction to receive income from such publication shall only extend to the publication's author(s). With regard to such restriction, in case of any inconsistency between the Creative Commons IGO 3.0 Attribution-NonCommercial-NoDerivatives license and these statements, the latter shall prevail.

Note that link provided above includes additional terms and conditions of the license.

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

BY NC ND

Abstract

Teachers are the most fundamental input of students' learning. For this reason, developing teaching skills is a policy priority for most governments around the world. We experimentally evaluate the effectiveness of "Let's All Learn to Read," a one-year professional development program that trained and coached teachers throughout the school year and provided them and their students with structured materials. Following a year of instruction by the trained teachers, students' literacy scores in treated schools grew by 0.386 of a standard deviation compared to students in the control group. These gains persisted through the second and third grades. We also show that an early intervention in first grade is more cost-effective at improving literacy skills than implementing remediation strategies in third grade.

JEL classifications: I20, H52, O15
Keywords: Teacher professional development, Early Interventions, Literacy, Colombia

Alvarez: Human Development, The World Bank (halvarezm@worldbank.org). Berlinski: Research Department, Inter-American Development Bank (samuelb@iadb.org). Busso: Research Department, Inter-American Development Bank (mbusso@iadb.org). Martinez-Correa: Research Department, Inter-American Development Bank (julianma@iadb.org). We thank Fundacion Luker and the Secretaria de Educacion Publica of Manizales for their support in implementing the evaluation. In particular, to Santiago Isaza, Maria Camila Arango, Monica Bedoya, Judith Estella, and Gloria de los Rios. The pedagogical material was developed and revised over time by Alejandra Mielke, Eira Cotto, Angela Marquez and Mauricio Duque. We thank Anna Koh for excellent research assistance. We thank the Korea Poverty Reduction Fund of the Inter-American Development Bank for providing funding which supported parts of this study. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

1 Introduction

More than 125 million children worldwide struggle to acquire basic literacy skills in early grades of schooling (Filmer et al. 2018). A key input for student learning is the quality of teaching (Chetty et al. 2014, Hanushek & Rivkin 2010). It is therefore not surprising that improving teachers' skills has long been a central concern of governments around the world. A popular strategy for achieving that goal is to offer in-service professional development to active teachers.[1] A frequently used alternative to promote the acquisition of literacy skills is to implement remediation programs to help students that struggle to achieve the minimum level of proficiency expected for their grade. In this paper, we evaluate the effectiveness of an in-service teacher professional development program and contrast its cost-effectiveness with a remediation program implemented in the same setting.

We experimentally evaluate the program "Let's All Learn to Read": an easily scalable teacher development program which bundled several components. The intervention provided teachers with new pedagogical tools designed to enhance literacy outcomes for children in early grades. The teaching method underpinning the intervention had three important ingredients that Alvarez-Marinelli et al. (2021) found to be effective in a small-group tutorial remediation program: it was based on a phonetic approach[2], it was designed based on an active pedagogy, and it followed a curriculum that was carefully structured. The in-service professional development intervention bundled several components (which Popova et al. (2021) show to be associated with larger student literacy gains). Teachers received intense, in-person training for two weeks, followed by continuous, in-class support coaching visits throughout the school year. The initiative also incorporated the development and distribution of complementary pedagogical material including books for teachers, and workbooks and storybooks for students.

[1] In developed countries, for example, 94 percent of surveyed teachers reported having attended at least one professional development activity in the year prior to the survey (OECD 2019).

[2] The phonics approach has been shown to work in a variety of contexts (Machin et al. 2018, Hirata & e Oliveira 2019). Although there are some exceptions (Jacob 2017).

The experiment consisted of a randomized controlled trial involving 70 schools (and close to 2,100 students), with 18 schools assigned to the treatment group and 52 to the control group. The main outcomes of interest are scores on the standard Early Grade Reading Assessment (EGRA), which includes four subtasks: knowledge of letter sounds, reading of non-words, fluency of oral reading, and reading comprehension. We aggregate the outcomes of these subtasks to create an overall literacy score. We measure students' scores at four points in time: at the end of first grade, the beginning of second grade, the end of second grade, and the beginning of third grade. This allows us to determine whether the potential learning gains varied over time.

Our paper relates to a large literature that assesses whether in-service teacher professional development programs are effective at improving students learning outcomes in developing countries. Some studies found no impacts on student outcomes (Loyalka et al. 2019, Zhang et al. 2013). Others show that the evaluated interventions were able to produce the intended changes in teacher practices but failed to translate them into higher student achievement (Berlinski & Busso 2017, Carneiro et al. 2022)–underscoring the importance of evaluating programs before scaling them up. A third group of papers found that teacher professional development programs that "bundle" similar components to those that were part of the program "Let's All Learn to Read" were effective at improving literacy outcomes in Kenya (Lucas et al. 2014, Piper et al. 2018), Uganda (Lucas et al. 2014, Kerwin & Thornton 2021)) and South Africa (Cilliers et al. (2020)).

Our study makes three contributions to this literature. First, we add new evidence that shows that "Let's All Learn to Read" led to an improvement of 0.386 of a standard deviation in the overall literacy proficiency score for students in treated schools at the end of the first grade. We attempt to indirectly shed some light on possible mechanisms behind our main results. Students experienced no gains in scores for mathematics – a subject not directly targeted by the program but one that could have been indirectly impacted had teachers applied components (such as active pedagogy) of the training program across subjects. We find that the treatment effects were stronger for teachers leading larger classrooms. We interpret this finding as indirect evidence that some classroom-management skills played a role in the treatment effects. We find that treatment effects were homogeneous across students of different characteristics such as gender, socioeconomic status, and initial levels of literacy skills. The equal level of gains experienced by students who initially had the lowest reading skills may reflect components of the program that gave teachers tools to promptly identify and help students who struggled to read.

A second contribution of this paper is to show that the positive effects of the in-service teacher professional development program did not fade out. In second and third grade, students exposed to treated teachers during first grade had literacy scores that were 0.3-0.4 standard in second and third grades deviations higher than students in the control group. These gains were observed even when second and third-grade teachers were not part of the teacher development program (which was exclusively designed for first grade literacy).

A third contribution of this paper is to show that an early intervention is more cost-effective than a later remediation. Early interventions are predicated on the fact that they achieve greater gains than comparable remedial interventions later in life and that they also save money in the future in terms of compensatory interventions (Carneiro & Heckman 2003). However, we do not often know what these remediation programs would look like, what their benefits are, and how much they will cost, which makes these comparisons speculative at best. This is not our case. The professional development program offered to first-grade teachers evaluated in this paper was built on the shoulders of a successful third-grade literacy remedial intervention. Moreover, both interventions were implemented in the same setting. This allows us to put on equal footing both interventions in order to compare their relative cost-effectiveness. Alvarez-Marinelli et al. (2021) show that a small-group remediation program led to 0.18σ learning gain per USD 100 spent. The teacher professional development intervention achieved a learning a gain of 1.03σ per USD 100 spent.

2 Intervention and Research Design

In 2018, the Secretary of Education of the Municipality of Manizales in Colombia, in partnership with a local NGO (Fundacion Luker), implemented an intervention ("Let's All Learn to Read") aimed at improving reading fluency of first-grade students in public schools by using teacher training and the development of pedagogical materials.[3] Seventy schools (and close to 2,100 students) participated in the experiment.[4] We created blocks of four schools of similar levels of enrollment, and we randomized schools within these blocks. Of the 70 schools, 18 were assigned to the treatment group and 52 to the control group.[5]

The program provided teachers with new pedagogical tools designed to better teach children to read from an early age. It had three important features that had underpinned a remediation program proven to be effective in the same setting (Alvarez-Marinelli et al. 2021). First, these tools focused on a phonetic approach, which targets the development and consolidation of key, foundational reading and writing skills, including phonological awareness, the alphabetic principle, the acquisition of new vocabulary, oral comprehension, reading comprehension, and the writing of letters, words and sentences.[6] Second, the approach was designed based on an active pedagogy. It encouraged children to make connections between printed and spoken words, the sound of alphabetic letters, and the rhythmic patterns of language, among other features. The program required the child to exercise and apply these skills in different situations and scenarios in an effort to achieve long-lasting learning. Third, the curriculum was carefully scaffolded and structured. Each week, teachers worked with students on a phoneme (i.e., a letter sound) lesson.[7] A phonological billboard presented each

[3] In 2021 the program was awarded a Qatar Foundation WISE Award, recognizing the project as an innovative undertaking that positively contributes to education and society.

[4] There were initially 71 schools randomized to treatment and control at the end of the 2017 school year. One school randomized to the control group did not open in 2018. This left a total of 70 schools that participated in the experiment.

[5] In 2019 an additional 18 schools received the treatment, and in 2020 the program was extended to all primary schools in Manizales.

[6] For more details on the phonetic approach see, for example, Foorman & Torgesen (2001) and NAEP (2000).

[7] The order of phonemes was based on Dehaene (2015).

phoneme with a character represented by an animal and short texts to develop comprehension strategies.[8] Classroom sessions were part of the normal school day. They lasted for one hour each day during 40 weeks and occurred during regular school hours. All students in the class were exposed to the new pedagogical approach. The program suggested that teachers use frequent evaluations so that they could intervene promptly to better help those students who were having difficulties. Teachers delivered remediation exercises aimed at ensuring a minimum level of reading skills.

The intervention also included the development of pedagogical materials. These included guidebooks for the teachers, and workbooks and storybooks for students. Students' workbooks contained exercises for practicing letters in word contexts, and for tracing and writing words and sentences from the letters learned. The teachers used the storybooks for read-aloud sessions with students.[9]A key element of the intervention was supporting teachers on the use of the new pedagogical approach. To that end, teachers received intense (in-person) training by external experts during five days throughout the academic year. In addition, they received regular support during the school year; every week, trained tutors made one-hour visits to help teachers by providing feedback and modeling teaching techniques.[10]

Teachers in schools randomized to the control group continued their teaching practices as before. During regular school hours students receive instruction according to the primary school curriculum, which includes four main academic subjects: Spanish, mathematics, natural sciences, and social sciences. In early grades, academic subjects are all taught by the same teacher. Although there are national guidelines regarding what children should achieve, schools and teachers are free to choose which pedagogical approaches and classroom strategies they use, and how much time they allocate to different subjects (MEN 2016). In general, regular class teaching of literacy in Colombia incorporates a hodgepodge of approaches. In

[8] For example, work on the "f" phoneme used a seal ("foca" in Spanish) named Fernanda, who was shown smiling to prompt a discussion about happiness ("felicidad" in Spanish).

[9] All materials pertaining to this intervention can be found at here.

[10] To illustrate the dynamics of a methodology and the proper use of materials, the tutor would demonstrate a given exercise for the students and the teacher.

Colombia, teaching children to read typically combines a "whole language" method with some syllabic components, rather than taking a phonics approach.[11] Teachers in schools randomized to the control group did not know that they had been randomized out of any intervention.

3 Data and Empirical Strategy

Our main outcomes of interest are measures of language development using the Early Grade Reading Assessment (EGRA), a research-based collection of individual subtasks that measure some of the foundational skills needed for reading acquisition in alphabetic languages (Dubeck & Gove (2015); p. 317).[12] Specifically, we analyze the following EGRA subtasks: i) knowledge of letter sounds (requiring students to sound the letter), ii) reading of non-words (requiring students to string letter sounds in words that do not have any meaning but follow a common orthographic structure), iii) fluency of oral reading (requiring students to read aloud a paragraph either recognizing words by sight or reading phonemes), and iv) reading comprehension (requiring students to respond to questions regarding the content of the paragraph read for the previous subtask).[13] We aggregate these outcomes into an overall literacy score, the sum of correct answers across all subtasks. All outcomes are standardized by the control mean and standard deviation.

Although our prespecified, main outcomes focus on reading skills, we also used the Early Grade Math Assessment (EGMA) to assess early grade mathematical competence. The math student achievement measures include simple additions and subtraction problems,

[11] There is a large literature that debates the benefits of using the "whole language" approach or a phonics approach for early literacy (Soler 2016). The pendulum has now swung in favor of phonics (NAEP 2000).

[12] EGRA was designed by RTI-International (2009) under the auspices of the U.S. Agency for International Development (USAID) and the World Bank. This open-source assessment tool has been applied in more than 65 countries for countrywide assessments and program evaluations (Dubeck & Gove (2015)). Children are allowed one minute to complete each subtask; if a child is unable to finish the subtask in that time, she moves to the next subtask. See Alvarez-Marinelli et al. (2021) for details on the metrics used.

[13] These tests were administered orally by trained enumerators in one-on-one sessions with a child, using a tablet. The application of the tests took on average fewer than 20 minutes per student. In the Data Appendix we report the test items administered in each evaluation period. Alvarez-Marinelli et al. (2021) find that the tests, when administered in a similar setting, have good psychometric properties.

comparison of natural numbers, sequencing or ordering of natural numbers, completing simple equations, and an aggregate math score (the average of the correct responses for the five math tasks).

We measure these outcomes at four points in time: at the end of first grade, the beginning of second grade, the end of second grade, and the beginning of third grade. We pool this information, and we estimate the following model:

$$Y_{isct} = \alpha + \theta T_s + \mu_c + \gamma_t + \epsilon_{isct} \tag{1}$$

where Y_{isct} is an outcome for student i attending school s of randomization strata c measured at time t. μ_c and γ_t are strata and time fixed effects. T_s is an indicator variable equal to one if the student was enrolled in a school s that was randomized to receive the teacher professional development intervention. Our parameter of interest, θ, captures the average intention-to-treat effect. Standard errors are clustered at the school level, the unit of randomization.

To separate short- and medium-term effects we also estimate:

$$Y_{isct} = \alpha + \sum_{t=1}^{4} (\theta_t \times P_t \times T_s) + \mu_c + \gamma_t + \epsilon_{isct} \tag{2}$$

where we interact the treatment indicator variable T_s with indicator variables for each time t ($t = 1$ for the end of first grade, $t = 2$ for the beginning of second grade, $t = 3$ for the end of second grade, and $t = 4$ for the beginning of third grade) to estimate θ_t (i.e., treatment effects for each time t).

4 Results

4.1 Balance and Attrition

We find no systematic differences between students in schools that were randomized to treatment and control groups. We cannot reject the null hypothesis that the average school-level characteristics at the beginning of the school year is equal in both groups. The average class size is 21 and about a quarter of the schools are rural. The average socio-demographic characteristics of students at the beginning of first grade is also similar in treated and control schools. Students in the experiment were, on average, 6 years old. Almost half of the students were girls. One-third belonged to households defined as being in the lowest socioeconomic-status group.[14] We have a set of measures designed by local experts to capture elementary-school readiness. Although we have this information for only a sub-sample of students, response rates are balanced across children with different treatment statuses. Students in treated and control schools seem to have similar levels of literacy and socio-affective skills. We find differences between the groups only in terms of the measure of students' motor skills. Overall, we take these results as confirmation that randomization produced treatment and control groups with similar pre-treatment characteristics.[15] Finally, we find no evidence of differential attrition between treatment and control schools. Specifically, we are not able to reject at the 5 percent level the null hypothesis of equality of attrition in all time horizons.[16]

4.2 Impact on Students' Literacy

Table 1 shows the intention-to-treat estimates from training teachers about how to adopt the new teaching approach and how to use new pedagogical materials for students' language development. Panel A presents the pooled treatment effects estimates for all reading subtasks and the composite literacy index. Column 1 shows that as a result of the intervention the number of letters correctly sounded by eligible students in treated schools improved by 0.479 of a standard deviation. Additionally, column 2 shows that treated students' reading of

[14] The System of Identification of Potential Beneficiaries of Social Programs (SISBEN) assigns a six-value socioeconomic status to households as follows: 1 (very low income), 2 (low income), 3 (medium-low income), 4 (middle class), 5 (upper-middle class), and 6 (upper class). We define low socioeconomic status as those students from households classified into either the first or second group.

[15] See Panels A-C of Appendix Table A.1 for details.

[16] See Panel D of Appendix Table A.1, which shows the proportion of students attending eligible schools at baseline that did not take an exam later on. For the results for the beginning of second grade, we do reject the null of equality of attrition at a 10 percent level. At the beginning of grade three the COVID-19 pandemic started to unfold. This meant that data were not collected in 36 schools. The table reports the proportion of students who took the tests in those schools that were open. If we were to consider all schools, the attrition rate would increase to 49.5 percent, and the rate would still be balanced between treatment and control schools (p-value=0.57).

non-words scores were 0.293 of a standard deviation higher than those of students in control schools.

Table 1: Treatment Effects on Literacy

	Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading comprehension	Literacy score
	(1)	(2)	(3)	(4)	(5)
Panel A: Pooled					
T	0.479***	0.293***	0.335**	0.179**	0.372***
	[0.080]	[0.100]	[0.139]	[0.088]	[0.134]
Panel B: By time horizon					
T x End G1	0.425***	0.254*	0.360	0.147	0.386**
	[0.101]	[0.144]	[0.224]	[0.164]	[0.169]
T x Beg. G2	0.579***	0.321***	0.384***	0.262***	0.399***
	[0.098]	[0.097]	[0.131]	[0.096]	[0.142]
T x End G2	0.423***	0.276**	0.273**	0.125	0.320**
	[0.087]	[0.110]	[0.127]	[0.085]	[0.138]
T x Beg. G3	0.495***	0.338***	0.319**	0.184*	0.388***
	[0.090]	[0.078]	[0.127]	[0.103]	[0.121]
Observations	6734	6734	6734	6734	6734
p-value End G2 <= End G1	0.506	0.388	0.730	0.566	0.800
p-value Beg. G3 <= End G1	0.191	0.271	0.575	0.429	0.494
Control mean at end G1	12.63	16.29	30.65	3.289	62.85
Control s.d. at end G1	13.60	12.45	22.33	2.723	43.81

Note: Panel A shows the results the results of estimating equation (1) for each column outcome. Each column shows the coefficients θ_h of equation (1), that is, the estimated treatment effects for all time horizons. Panel B shows the results the results of estimating equation (2) for each column outcome. Each column shows the coefficients θ_h of equation (2), that is, the estimated treatment effects at different time horizons for each outcome of interest. The rows labeled "p-value End G2 <= End G1" and "p-value Beg. G3 <= End G1" show the p-value of a test $H_0 : \theta_2 <= \theta_1$ and $H_0 : \theta_3 <= \theta_1$, respectively. Control mean and S.D. correspond to the number of correct answers. All models include strata, and date fixed effects. Standard errors, shown in squared brackets, are clustered at the school level (the unit of randomization). * significant at 10%; ** significant at 5%; *** significant at 1%.

Column 3 of Table 1 shows the effect of the treatment on the fluency of reading aloud a text at the appropriate grade level. We find that students in treated schools gained 0.335 of a standard deviation over their peers in control schools. Moreover, column 4 shows that students in the treated schools improved by 0.179 of a standard deviation in the subtask that focused on the comprehension of the text that the children read aloud. Finally, column 5 shows estimates of the impact on overall literacy by adding the number of correct answers in all subtasks; we find gains of 0.372 of a standard deviation for students in treated schools. These treatment effects are large. For comparison, (Cilliers et al. 2020) find that providing early-grade language teachers with in-class coaching for two years led to gains in student achievement tests of about 0.24 of a standard deviation. He also finds that centralized training of similar teachers produced learning gains of 0.12 of a standard deviation.

The size of the treatment effects can be explained by the combination of two factors: the nature of the content of the program, and fidelity to its implementation. First, the design of the teacher professional development intervention included features that had previously been identified as effective (Popova et al. 2021). These were subject-specific pedagogy (in our case, a phonetic approach); centralized, face-to-face training plus continuous, in-class support with tutors visiting teachers every week; lesson enactment in the training (in our case, exercises demonstrated by tutors in the accompanying visits); complementary materials in the form of tailored storybooks and workbooks for both teachers and students; and carefully structured lessons with specific guidelines for working each week.

In addition, our intervention was implemented with high fidelity. Tutors regularly observed classes and collected qualitative information. They reported three main observations. First, teachers and students were highly engaged with the new curriculum. Teachers carried out all the materials' proposed activities, including the homework and the suggested formative evaluations. The oral reading done by the teacher had good rhythm, intonation, and good use of orthographic signs. These characteristics were important to properly model reading to the students. As a result, students showed interest in reading, and teachers developed confidence about the program's effectiveness. Second, tutors emphasized that teaching was fine-tuned to students' needs by providing remediation for those who were lagging behind. Third, tutors coached teachers to improve teaching. Some teachers had difficulties in the delivery of the lessons. For example, some teachers tended to ignore the pre-reading activities, or they slightly changed the way they asked reading comprehension questions. In these situations, tutors worked with teachers to better implement the new curriculum.

Panel B of Table 1 presents dynamic treatment-effect estimates following equation (2). We find that gains are very persistent over time. Results remain positive and large for almost all outcomes at all time horizons.[17] Point estimates for second and third grades are at least of the same size as those at the end of first grade, suggesting that the impact did not fade out. At the bottom of the table we present one-sided p-values of the null hypothesis that the coefficients in second or third grade are smaller than those observed at the end of first grade. In all cases, we cannot reject the null hypothesis of equal treatment effects. Our intervention was designed for and ended in first grade: the new pedagogical approach and materials were specifically designed for this grade level. In second and third grades students that participated in the experiment were taught by different teachers who did not receive any teacher professional development (beyond what is offered routinely by the Secretary of Education of Manizales). This suggests that our early intervention had long-lasting effects on children's literacy outcomes.

Robustness. Figure 1 presents various robustness checks for the treatment effect estimates of each literacy outcome. First, we estimate treatment effects at end of first grade for a second round of the experiment run in 2019.[18] We find positive results for all literacy subtasks (although the results are not always statistically significant at standard levels). For the composite literacy score, the treatment effect is statistically significant and close to that of the first cohort.[19]

Because there is some minor evidence of attrition, we estimate upper and lower bounds for our main treatment effects following Lee (2009). We find that the lower bound of the treatment effect remains positive and statically significant for all outcomes. We also found small differences in motor skills measured at the end of kindergarten between children in treated and control groups. Figure 1 shows that results remain unchanged if we add end-of-

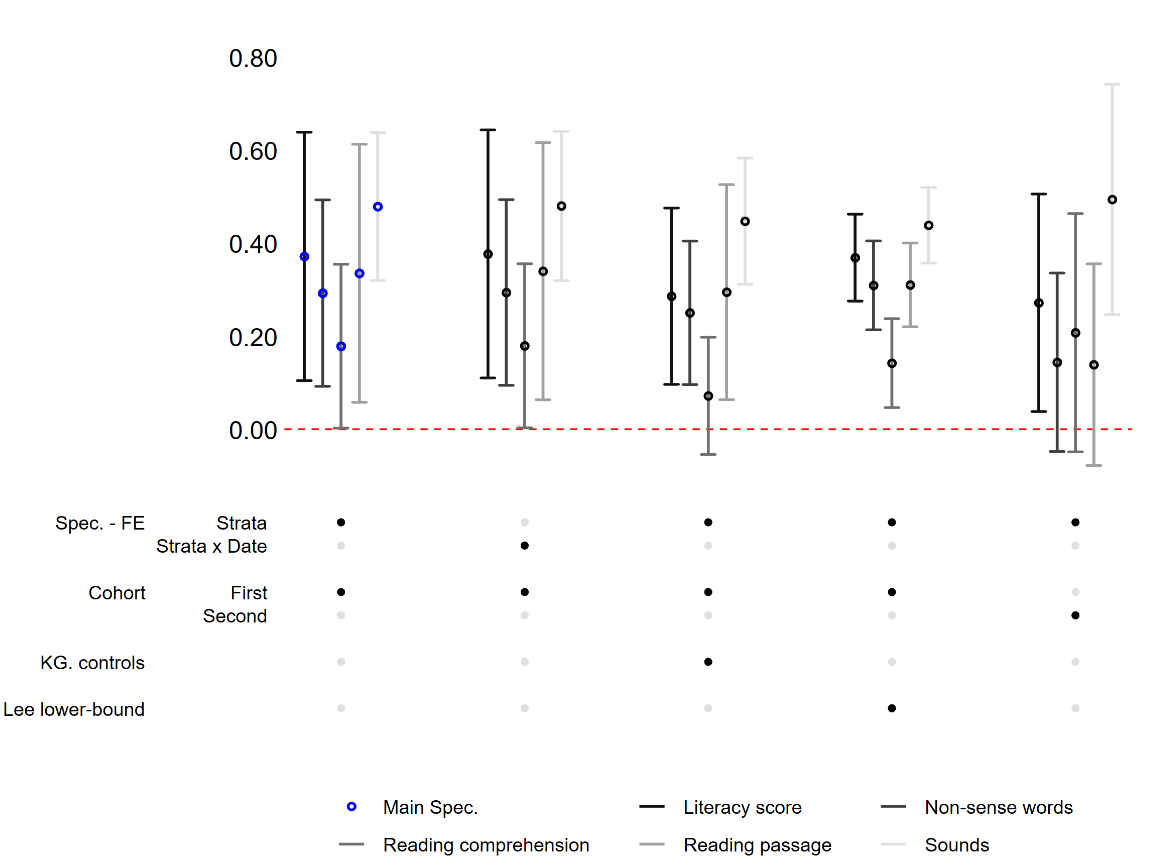
[17] Reading comprehension is the only outcome that presents a more irregular pattern; it is larger at the beginning of second and third grades.

[18] For this round, 18 more schools were added to the treatment group. We excluded this second cohort of the experiment from our main analysis for two reasons. First, some schools did not open a first grade in 2019, and, as a result, those first-grade students were reallocated to nearby schools. This affected somewhat the validity of the experiment in this second cohort. Second, we are unable to analyze long-term outcomes for this second cohort because schools were affected by closures during the COVID-19 pandemic.

[19] Appendix C presents the balance table, attrition, and full set of results for the second round of the experiment.

kindergarten characteristics as control variables in estimating equation (1). Finally, equation (1) imposes the same strata fixed effect for all time horizons. We include strata × date fixed effects, and we find that results do not change.

Figure 1: Robustness



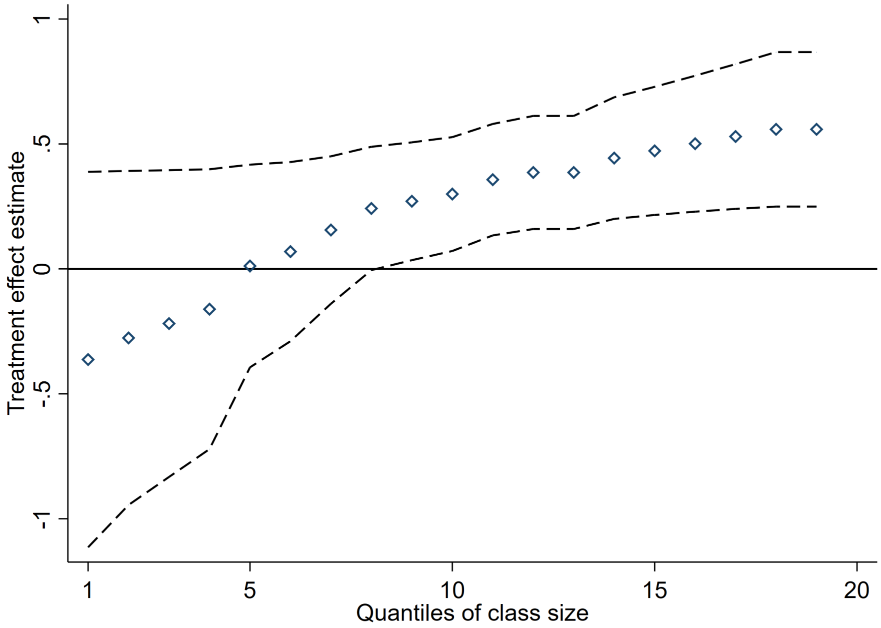
Note: The figure plots the treatment coefficients (and 90% and 95% confidence intervals) for each literacy outcome for the main estimation and different robustness checks, including strata × date fixed effects, controlling for end-of-kindergarten characteristics, Lee (2009) lower bounds, and results for the second round of the experiment at end of first grade. Each combination is represented by black/white dots at the bottom of each subfigure.

4.3 The Role of Classroom Management in Explaining the Results

The new teaching curriculum embedded three components: a phonetic approach to literacy teaching, active pedagogy, and structured literacy lessons that facilitated classroom management. Our research design does not allow us to disentangle the relative contribution of each component to the overall effectiveness of the program. However, we can indirectly analyze the role that classroom management played in our setting. We provide two pieces of evidence for this.

First, we conjecture that better classroom management partly explains the positive treatment effects on literacy. Literacy lessons were carefully structured for the school year. Teachers followed the intervention materials with specific learning objectives and exercises designed for each lesson. This structure could have helped teachers better manage large classrooms, which usually constrain students' learning.[20]

Figure 2: Treatment Effect on Literacy Scores by Class Size



Note: The figure shows linear predictions and 95% confidence intervals of the treatment effect estimates on literacy scores. We compute this based on the treatment and the treatment-class size interaction coefficients of a regression analogous to equation (1). The regression additionally includes class size as a control variable. The standard error for estimates at each quantile v is constructed as $\sqrt{Var(\hat{\delta} + \hat{\beta}_Z \times \overline{Z}_v)}$, where $\overline{Z}_v$ is the mean of class size in quantile v.

We explore whether treatment effect estimates vary with class size. We estimate equation (1) adding an interaction term between the treatment and class size (as well as class size as a control variable). Figure 2 plots the linear prediction of the treatment effects on the aggregate literacy score for students who attended classes in classrooms of different sizes. We find that treatment effect estimates are statistically significant only for large classrooms

[20] For instance, Urquiola (2006) finds that reducing class size by (on average) nine students increases test scores from 0.16 of a standard deviation to 0.30 of a standard deviation.

(more than 25 students). These results are consistent with those of Cilliers et al. (2020), who finds that training and coaching approaches have larger impacts in classrooms with close to 40 students per class.

Table 2: Treatment Effects on Other Outcomes

	Additions	Subtractions	Comparing numbers	Ordering numbers	Equations	Math score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pooled						
T	0.010	–0.064	0.055	0.033	–0.037	–0.011
	[0.086]	[0.083]	[0.065]	[0.083]	[0.058]	[0.106]
Panel B: By time horizon						
T x End G1	–0.037	–0.197	0.107	0.137	–0.069	–0.045
	[0.147]	[0.133]	[0.109]	[0.130]	[0.126]	[0.152]
T x Beg. G2	0.027	0.013	–0.042	0.046	0.034	0.015
	[0.087]	[0.114]	[0.070]	[0.099]	[0.098]	[0.127]
T x End G2	0.064	0.015	0.062	–0.062	–0.080	0.009
	[0.106]	[0.103]	[0.067]	[0.090]	[0.058]	[0.113]
T x Beg. G3	–0.030	–0.102	0.114*	–0.005	–0.031	–0.030
	[0.083]	[0.075]	[0.064]	[0.118]	[0.063]	[0.093]
Observations	6734	6734	6734	6734	6734	6734
Control mean at endline	7.894	6.062	7.017	3.347	3.785	28.10
Control s.d. at endline	4.044	3.806	2.830	1.914	2.442	11.84

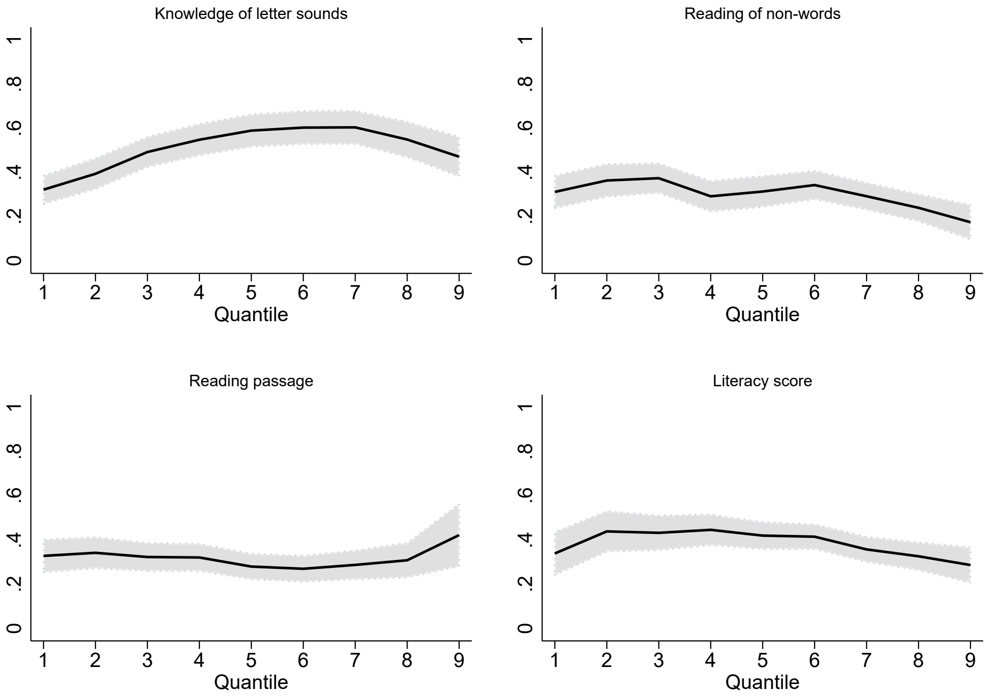
Note: Panel A shows the results the results of estimating equation (1) for each column outcome. Each column shows the coefficients θ_h of equation (1), that is, the estimated treatment effects for all time horizons. Panel B shows the results the results of estimating equation (2) for each column outcome. Each column shows the coefficients θ_h of equation (2), that is, the estimated treatment effects at different time horizons for each outcome of interest. The rows labeled "p-value End G2 <= End G1" and "p-value Beg. G3 <= End G1" show the p-value of a test $H_0 : \theta_2 <= \theta_1$ and $H_0 : \theta_3 <= \theta_1$, respectively. Control mean and S.D. correspond to the number of correct answers. All models include strata and date fixed effects. Standard errors, shown in squared brackets, are clustered at the school level (the unit of randomization). * significant at 10%; ** significant at 5%; *** significant at 1%.

Second, this apparent improvement in classroom management seemed to be tightly linked to the fact that literacy lessons were highly structured. Because first-grade teachers provide instruction on both literacy and math, better classroom management could have potentially be transferred to math teaching. We test this in Table 2, which shows the results for six math outcomes – not directly targeted by the intervention. We find that the treatment effects were quantitatively small and not statistically significant in first, second, or third grade (with signs changing across the different outcomes). Thus, we surmise that seeking to improve students' math learning through teacher training might require a professional development program specifically designed to change how math is taught.

4.4 All Students Gained with the New Curriculum

Literacy gains were homogeneous among children of different literacy skills. We follow Firpo (2007) to estimate quantile treatment effects for all subtasks. Figure 3 shows that, even though the impact of the intervention was slightly heterogeneous on knowledge of letter sounds, the relationship was flat for the other subtasks as well as for the aggregate literacy score.

Figure 3: Quantile Treatment Effects



Note: Each panel shows the quantile treatment effects on each outcome of interest estimated following Firpo (2007). The reading comprehension outcome could not be estimated due to convergence problems.

This could be explained by the fact that the intervention improved teachers' ability to teach at the right level by providing continuous assessments to identify struggling readers and in-time remedial exercises to help them. Moreover, these remediation practices were encouraged by the tutors in their class visits. This result is similar to those found by Muralidharan et al. (2019). They show that an adaptive, computer-based intervention designed to teach at the right level by anchoring content and assessments to students' initial knowledge increased math scores by 0.37 of a standard deviation and language scores by 0.23 of a standard deviation; the results were homogeneous for students with different initial test scores.

We also study how the treatment effects vary according to students' sex and socioeconomic status. We estimate equation (1) for each subgroup and test the null hypothesis that the estimated coefficients in each subgroup are equal. We do not find differences in literacy gains between girls and boys. Even though the estimated treatment effects are larger for students with higher socioeconomic status (than for those in households of low socioeconomic status), we cannot reject equality of the estimated coefficients.[21]

5 The Benefits of Early Interventions

The professional development program offered to first-grade teachers evaluated in this paper was built on the shoulders of a successful third-grade literacy remedial intervention. Both interventions were experimentally evaluated in the same setting. Early interventions are predicated on the fact that they can achieve greater gains than comparable remedial interventions later in life and that they also save money in the future in terms of compensatory interventions (Carneiro & Heckman 2003). Our setting allows us to provide empirical evidence for these ideas.

Alvarez-Marinelli et al. (2021) study this remediation program piloted in 2015-2017 in the same universe of schools of the Municipality of Manizales that participated in the experiment analyzed in this paper. The remediation program consisted of a set of 40-minute structured tutorial sessions provided three times a week during the school day for up to 16 weeks. The sessions were conducted in small groups and followed a simple structure based on a phonics approach. The tutorials were led by tutors who where hired and trained to deliver the program. The program cost USD 89 per student and produced gains in literary of 0.27σ.

By comparison, the in-service teacher professional development intervention studied in this

[21] Results are reported in Appendix Table A.2.

paper led to average aggregate literacy learning gain of 0.37σ at a cost of USD 36 per student. These costs were mainly driven by wages of the personnel who trained and coached teachers, and by the opportunity costs of teachers' time, given that the general training was conducted out of regular school hours. Our calculations exclude the fixed costs of materials development (books for teachers, storybooks and workbooks with exercises for students) since the contribution to the cost per student tends to move toward zero as the program is scaled up.[22]

In our setting, early intervention was more cost-effective than later remediation. Alvarez-Marinelli et al. (2021) estimate a 0.18σ learning gain per USD 100 spent in literacy remediation of third-grade students of the same schools included in our setting. By contrast, the teacher professional development intervention achieved a learning gain of 1.03σ per USD 100 spent.[23,24] Furthermore, children were selected to receive the remedial literacy program if at the beginning of third grade they read fewer than 60 words per minute. One-third of students were deemed eligible for the program with this criterion. The gains from the teacher professional development intervention studied in this paper imply that only 22 percent of children would meet the criteria for remediation in third grade.[25] Therefore, the intervention saves about USD 10 per student.[26]

[22] Appendix Table A.3 breaks down the costs in each category.

[23] Evans & Yuan (2019) estimate that students in Colombia gain 1σ in reading proficiency in 4.8 to 9.3 years of schooling. These values would imply less conservative *annual* learning gains (between 0.11 and 0.21σ, compared to the 0.4σ used from Alvarez-Marinelli et al. (2021)).

[24] In a similar coaching program in South Africa, Cilliers et al. (2020) estimate a 0.57σ increase in literacy per USD 100 spent per student annually.

[25] Let S^0_{3i} be the number of words that student i reads fluently at the beginning of third grade. Let $W^+ = 0.319 \times \sigma_3 = 8.2$ be the average increase in the number of words fluently read due to the teacher professional development intervention. We can compute the relative change in the number of students who are eligible for remediation as $\frac{N^{E\to NE}}{N^E} = \frac{\sum_{i=1}^{N^E}[1(S^0_{3i}+W^+>60)]}{\sum_{i=1}^{N}[1(S^0_{3i}]} = \frac{482}{1467} \approx \frac{1}{3}$ (where N^E denotes the number of students who are eligible for remediation absent the first-grade teacher professional development intervention and $N^{E\to NE}$ are the number of students who, because of that intervention, no longer require remediation). Hence, $\frac{1}{3} - \frac{1}{3} \times \frac{1}{3} \approx 0.22$.

[26] Let N be the total number of students in a class. Before the in-service teacher professional development program was implemented remediation costed $89 \times N \times 0.33$ per student. With the professional development program in place, the cost of remediation would be $89 \times N \times 0.22$. Therefore, we save $89 \times N \times 0.11$ as a result of the professional development program.

6 Conclusion

We present the results of a professional development program that provided teachers with new pedagogical tools designed to enhance literacy outcomes for first-grade children. The "Let's All Learn to Read" program was a one-year intervention that consisted of in-person teacher training and continuous, in-class support throughout the year to achieve a correct implementation of the new phonetic approach and carefully structured curriculum. At the end of first grade, when the intervention finished, literacy proficiency scores of the treated teachers' students had improved by 0.386 of a standard deviation; the gains remained consistent throughout second and third grades. This teacher professional development program brings together many of the characteristics identified in the literature as being associated with larger student gains, and the program was implemented with very high fidelity. This combination seems to explain the large impacts, which in turn translate into a high degree of cost-effectiveness.

Early interventions are sometimes preferred to interventions later in life based on their relative learning gains and cost-effectiveness. We are able to compare the in-service teacher professional development program for first grade teachers with a remedial tutorial program for third-grade students in the same schools. We show that the professional development program for first-grade teachers was more cost-effective than the later remediation program (1.03σ vs. 0.18σ learning gain per USD 100 spent). We estimate that our early intervention can save about USD 10 per student in third-grade remediation.

The scaling-up of the program is already in motion. Currently, more than 700,000 public school students in Colombia and Panama are benefiting from the program. These numbers are expected to increase soon. Materials are being adapted for Portuguese in Brazil while Ecuador and Dominican Republic have also shown interest in adopting this teaching model. These recent developments, which incorporate effective program's attributes to scaled-up government-funded teacher development programs, suggest that the gap between evidence and policy can be reduced.

References

Alvarez-Marinelli, H., Berlinski, S. & Busso, M. (2021), 'Remedial education: Evidence from a sequence of experiments in colombia', *Journal of Human Resources* .

Berlinski, S. & Busso, M. (2017), 'Challenges in educational reform: An experiment on active learning in mathematics', *Economics Letters* **156**, 172–175.

Carneiro, P., Cruz-Aguayo, Y., Schady, N., Ruthy, I., Juan, P. & Sarah, S. (2022), 'When promising interventions fail: Personalized coaching for teachers in a middle-income country', *Journal of Public Economics Plus* .

Carneiro, P. M. & Heckman, J. J. (2003), 'Human capital policy', *National Bureau of Economic Research* (9495).

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014), 'Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood', *American Economic Review* **104**(9), 2633–79.

Cilliers, J., Fleisch, B., Prinsloo, C. & Taylor, S. (2020), 'How to improve teaching practice? an experimental comparison of centralized training and in-classroom coaching', *Journal of Human Resources* **55**(3), 926–962.

Dehaene, S. (2015), *Aprender a leer: de las ciencias cognitivas al aula*, Siglo XXI Editores.

Dubeck, M. M. & Gove, A. (2015), 'The early grade reading assessment (egra): Its theoretical foundation, purpose, and limitations', *International Journal of Educational Development* **40**, 315 – 322.

Evans, D. & Yuan, F. (2019), 'Equivalent years of schooling: A metric to communicate learning gains in concrete terms', *World Bank Policy Research Working Paper* (8752).

Filmer, D., Langthaler, M., Stehrer, R. & Vogel, T. (2018), 'Learning to realize education's promise', *World Development Report. The World Bank* .

Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**(1), 259–276.

Foorman, B. & Torgesen, J. (2001), 'Critical elements of classroom and small-group instruction promote reading success in all children.', *Learning Disabilities Research and Practice* **16**(4), 203–212.

Hanushek, E. A. & Rivkin, S. G. (2010), 'Generalizations about using value-added measures of teacher quality', *American Economic Review* **100**(2), 267–71.

Hirata, G. & e Oliveira, P. R. (2019), 'Lasting effects of promoting literacy – do when and how to learn matter?', *Education Economics* **27**(4), 339–357.

Jacob, B. (2017), 'When evidence is not enough: Findings from a randomized evaluation of evidence-based literacy instruction', *Labour Economics* **45**, 5 – 16.

Kerwin, J. T. & Thornton, R. L. (2021), 'Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures', *The Review of Economics and Statistics* **103**(2), 251–264.

Lee, D. S. (2009), 'Training, wages, and sample selection: Estimating sharp bounds on treatment effects', *The Review of Economic Studies* **76**(3), 1071–1102.

Loyalka, P., Popova, A., Li, G. & Shi, Z. (2019), 'Does teacher training actually work? evidence from a large-scale randomized evaluation of a national teacher training program', *American Economic Journal: Applied Economics* **11**(3), 128–54.

Lucas, A. M., McEwan, P. J., Ngware, M. & Oketch, M. (2014), 'Improving early-grade literacy in east africa: Experimental evidence from kenya and uganda', *Journal of Policy Analysis and Management* **33**(4), 950–976.

Machin, S., McNally, S. & Viarengo, M. (2018), 'Changing how literacy is taught: Evidence on synthetic phonics', *American Economic Journal: Economic Policy* **10**(2), 217–41.

MEN (2016), *Derechos Básicos de Aprendizaje: Lenguaje*, Ministerio de Educación Nacional.

Muralidharan, K., Singh, A. & Ganimian, A. J. (2019), 'Disrupting education? experimental evidence on technology-aided instruction in india', *American Economic Review* **109**(4), 1426–60.

NAEP (2000), 'National assessment of educational progress at grade 4', *National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.* .

OECD (2019), *TALIS 2018 Results (Volume I)*.

Piper, B., Zuilkowski, S. S., Dubeck, M., Jepkemei, E. & King, S. J. (2018), 'Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides', *World Development* **106**, 324–336.

Popova, A., Evans, D. K., Breeding, M. E. & Arancibia, V. (2021), 'Teacher Professional Development around the World: The Gap between Evidence and Practice', *The World Bank Research Observer* .

RTI-International (2009), 'Early grade reading assessment toolkit', *World Bank Working Paper, Office of Human Development* .

Soler, J. (2016), 'The politics of the teaching of reading', *Prospects* **46**(3), 423–433.

Urquiola, M. (2006), 'Identifying class size effects in developing countries: Evidence from rural bolivia', *The Review of Economics and Statistics* **88**(1), 171–177.

Zhang, L., Lai, F., Pang, X., Yi, H. & Rozelle, S. (2013), 'The impact of teacher training on teacher and student outcomes: evidence from a randomised experiment in beijing migrant schools', *Journal of development effectiveness* **5**(3), 339–358.

A Appendix Tables and Figures

Table A.1: Balance and Attrition

		Treatment	Control	p-value of $H0 : \beta_t = \beta_c$	Obs.
Panel A:	Avg. schools' characteristics				
	Rural school	0.33	0.23	0.18	70
	School size	284.11	329.08	0.33	70
	Class size	21.94	20.70	0.41	70
	Number of classrooms in Grade 1	1.44	1.33	0.28	70
	Literacy	0.64	0.64	0.88	70
	Math score	0.49	0.50	0.23	70
Panel B:	Avg. beg. G1 students' characteristics				
	Age	6.00	5.98	0.74	2227
	Female	0.46	0.47	0.88	2227
	No disability	0.95	0.96	0.42	2227
	Low socio-econ status	0.65	0.68	0.43	2227
Panel C:	Avg. end KG students' characteristics				
	Prob. response	0.60	0.62	0.56	2227
	Literacy	0.21	0.00	0.11	1361
	Motor	0.29	0.00	0.02	1361
	Socio-affective	–0.06	0.00	0.74	1361
Panel D:	Attrition				
	End Grade 1	0.22	0.22	0.91	2227
	Beg. Grade 2	0.23	0.18	0.05	2227
	End Grade 2	0.21	0.21	0.85	2227
	Beg. Grade 3	0.31	0.32	0.59	1565

Note: Panel A shows the average school-level characteristics. Literacy and math scores are expressed as average proportion of correct answers in each subtask. They are averages of school-year averages (third and fourth grades in years 2015-2018). Panel B shows the average beginning of first grade characteristics of eligible students. Panel C shows the average end of kindergarten characteristics of eligible students and the probability of response for these data. Literacy, motor and socio-affective characteristics are simple averages of indicator sub-items that compose each variable. We then standardized to the mean and standard deviation of the control group. Panel D shows the attrition rates at different time horizons of the eligible students observed at baseline. Columns labeled "Treatment" show the average for students in schools randomized to treatment. Columns labeled "Control" show the average for students in schools randomized to control. Column labeled "p-value of $H0 : \beta_t = \beta_c$" presents the p-value of the treatment indicator coefficient in a regression with each row variable as the outcome. All regressions include strata fixed effects, and standard errors are clustered at the school level.

Table A.2: Treatment Effect Heterogeneity

		Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading comprehension	Literacy score
		(1)	(2)	(3)	(4)	(5)
Student's sex	Male	0.474***	0.322***	0.394**	0.175*	0.410**
		[0.091]	[0.115]	[0.163]	[0.093]	[0.160]
	Female	0.484***	0.256**	0.273**	0.185*	0.340***
		[0.085]	[0.103]	[0.127]	[0.102]	[0.119]
	p-value of equal coeffs.	0.845	0.440	0.270	0.857	0.626
Socio-econ. status	High	0.434***	0.312**	0.379**	0.172*	0.432***
		[0.100]	[0.123]	[0.181]	[0.089]	[0.153]
	Low	0.471***	0.267**	0.294**	0.151	0.311**
		[0.087]	[0.107]	[0.133]	[0.096]	[0.139]
	p-value of equal coeffs.	0.854	0.607	0.460	0.621	0.292

Note: Each panel shows estimates of treatment coefficient of equation (1) estimated separately for two different groups of control and treated students. The p-values of equal coefficients correspond to the interaction between treatment and the heterogeneity variable in a separate regression. Homogeneous classroom is an indicator variable that takes value one for classrooms with a s.d of an index of comparable subtasks measured at baseline below the median. All models include cohort, year, and strata fixed effects. The heterogeneity variable is also included as control. Standard errors, shown in squared brackets, are clustered at the school-level (the unit of randomization). * significant at 10%; **significant at 5%; *** significant at 1%.

Table A.3: Cost-Effectiveness

Cost per Teacher		**791**
Teacher Development		
Initial Training	8 hours a day for 5 days. 15 Teachers per training session	25
Teacher's time	8 hours a day for 5 days	380
Follow-up	1 hour per week per teacher (for 40 weeks)	380
Materials	2 books per teacher	6
Cost per Student		**36**
Teacher's cost	Class size = 24	33
Materials	1 book per student	3

Note: Parameters: Teachers per tutor session 15; Exchange rate 3,000 Colombian pesos per dollar of 2016; Caverage class size 24; Total number of students in tutorials 609; Hourly wage 4.167; Book costs 3.

B Data Appendix

B.1 Data Sources

This paper relies on three sources of information. First, at the end of the academic year, we administered language and mathematics tests to first grade students randomized public schools to measure the impact of the intervention. We collected same information in these schools at the beginning and the end of Grade 2 and at the beginning of Grade 3 using the same instrument as well. The former data consists of this test score information administered at each point in time. Details about the test design, content, scoring, and administration are presented in Section B.3.

At the time of baseline, we collected data on first grader students' developmental assessment which measures a child's attainment of motor, cognitive, communicative and socioemotional skills. In Colombia, this developmental test (Predictores de aprendizaje) is administered when a child is in kindergarten. In addition to this baseline data, we also collected administrative school records from the Integrated Enrollment System (Sistema Integrado de Matrícula, SIMAT), the national database for the registration of students in public education in Colombia. The latter provides information on students' age, gender, socio-economic status, and whether or not students change schools or repeat grades over time.

B.2 Sample Sizes

In Table B.1 we show the number of schools, classrooms and students that participated in the experiment. 70 public schools participated in 2018, 18 randomized to treatment and 52 to the control group. These implied a total of 26 and 69 classrooms for control and treatment, respectively. A total of 1,271 students were eligible for the intervention. It should be noted that when we measure the outcomes during Grade 2 we also follow up any retained students in first grade who initially were in our sample either in treatment or control groups but have not been promoted to Grade 2 during the next academic year.

Table B.1: Sample Sizes

	Treatment	Control	Total
Schools	18	52	70
Classrooms	26	69	95
Students	609	1562	2171

B.3 Instruments and Test Scores

As described in Section 2 our measures of student learning are captured by EGRA and EGMA tests (Early Grade Reading/Mathematics Assessment). The test contains four subtasks of literacy and five subtasks of math: knowledge of letter sounds, reading of non-words, fluency of oral reading, reading comprehension, addition, subtraction, comparing numbers, ordering numbers and equations. In the first component, children are asked to recognize the letter sounds and the next subtasks ask children to read invented words. After that, a simple passage is given to students and they are asked to read it aloud and answer several questions about it. The last five subtasks of math involve students solving math operation of one- and two- digit numbers to measure their math knowledge. We then combine all these subtasks scales into an aggregate score that measures literacy and math knowledge: literacy score and math score.

Each subtask was scored separately by counting the number of correct answers. All these raw scores were standardized within grade-subtask by the mean and standard deviation observed in the control group at the corresponding point of measurement. We also normalized the aggregate literacy and math scores, which are the proportion of correct answers in all subtasks.

B.4 Test Subtasks Over Time

Table B.2 presents how the test scales vary from each subtask in terms of item construction. Since it is common for an existing EGRA instrument to be modified into one or more parallel versions, the instrument we administered at each point in time has been modified by re-randomizing the items with grade-level equivalents in the first three literacy subtasks and equation subtask. Even though some minor scaling differences may exist, the outcomes are comparable across the different grades because the items have been modified in order to be as similar as possible in terms of length and difficulty.

Table B.2: Scales by Date

	End G1	Beg. G2	End G2	Beg. G3
Scales:				
Knowledge of Letter Sounds	1	2	3	3
Reading Passage	1	2	2	4
Reading Comprehension	1	2	2	4
Equations	1	2	2	3
Reading of Non-words, Comparing Numbers, Ordering Numbers, Additions, Subtractions	Always the same scales in all dates			

Tables B.3 and B.4 presents all the items for all the subtasks and tests. Column 2 shows the different composition of items for the Knowledge of Letter Sounds. Letters of the alphabet are distributed randomly and evenly among the upper- and lowercase letters, ten letters to a line. As mentioned before, we have three different scales for this subtask but with equivalent test items in terms of difficulty for each one. We look next at the reading fluency which is a one-paragraph passage that contains same sentence structures and complexity. Note that these two subtasks, as well as reading comprehension, are not constructed with a constant number of items across instruments. For example, the scale of Knowlege of Letter Sounds for the first graders contains 100 alphabets while the scale for the second graders only has 50 alphabets. On the contrary, the number of items of Reading Fluency is larger for second and third graders than that of first graders. However, despite these differences, the scores are comparable since we created a composite score by standardizing them for each date to place students on the same scale.

Table B.3: Language Test items

ID Scale	Knowledge of Letter Sounds	Reading of Non-words	Reading passage	Reading Comprehension
1	V l e m S y f ñ L N l K T D n T a d z w r ch z m U e j G X u g R B Q l f j Z s r B n C B p Y F c a E y s a Ll N P E M v Y O p t x n N k P c Z A D e d y x f b F j r r u v A Ch t G T b S l g m i l L L o q	lete quibe bofa mise garo cafa Celu bede lura mesi lluno Rite duso jata fica luma Alti lufa frate dulte ledo Fosu gesa lemo golpa bosa rale flano trabu bulo pluva arcu cince llusia firta onti zaca queno bana juru foba lise vodo tuzi listu quira cuto ganco rafo duba	María tiene una gata. A la gata le gusta jugar. Un día María no encontró a su gata. María y su mamá la buscaron por toda la casa. De pronto oyeron ”miau, miau”. Los maullidos eran suaves. Venían de debajo de la cama. María y su mamá encontraron a la gata y dos gatitos. La gata de María tuvo gatitos. La mamá de María le dijo: Yo también tendré un bebé. vas a tener un hermanito. María sonrió y se fue corriendo a la casa de su amiga Lorena. Al llegar le dijo a Lorena: ”Vengo a contarte grandes noticias”.	¿Quién es la dueña de la gata? ¿Qué le gusta hacer a la gata? ¿Por qué está preocupada María? ¿Por qué crees que la gata se escondió? ¿De dónde salían los maullidos? ¿Por qué crees que eran suaves los maullidos? ¿Cuántos gatitos tuvo la gata de María? ¿Qué le dijo la mamá a María? ¿Para dónde se fue María tan apurada? ¿Qué noticias le va a dar María a Lorena?
2	M d r O E F i u p S A n j T b e f r G L m R D E y O a g s Z f V N I b U P R L M S v O A d T i N a e		La Gallina y el Cienpiés se pusieron a jugar al fútbol para ver quién era el mejor jugador. Se fueron a la cancha y comenzaron a jugar. La Gallina era rápida, pero el Cienpiés fue más rápido. La Gallina pateó lejos, pero el Cienpiés pateó más lejos. La Gallina comenzó a enojarse. La Gallina anotó un solo gol en todo el juego. El Cienpiés con sus múltiples patas atrapó muchas pelotas. El Cienpiés anotó cinco goles en total. La Gallina estaba furiosa porque perdió. El Cienpiés se echó a reír. Después del partido la Gallina estaba tan enojada que abrió su pequeño pico y se tragó el Cienpiés de un solo bocado. De camino hacia su casa, la Gallina se encontró con la madre del Cienpiés quien le preguntó por su hijo.	¿Qué se pusieron a jugar la gallina y el cienpiés? ¿A dónde fueron a jugar fútbol? ¿Quién fue más rápido? ¿Quién pateó más lejos? ¿Quién anotó un solo gol?
3	c V N I k U P x L Q A n j T b e f r W L m R D E y O a g s Z c V N I k U P x L Q S Ñ O A d T i N a E			
4			El abuelo tomaba café. Era una tarde lluviosa. Recordaba cuando era niño. El abuelo contó , como era la siembrea de café. Él vivía en un pueblo. El pueblo era grande. El pueblo se llamaba Neira. Al regresar de la escuela, ayudaba a su papá a sembrar café. Le pregunté: como se siembra el café? El abuelo dijo: el café es una planta. Empieza siendo una semilla. Esta crece y se convierte en cafeto. El cafeto da un fruto rojo llamado cereza. Al madurar, se corta. Luego se seca al sol en grandes patios. Despúés se tuesta y se muele. El café se empaca y se vende. Esto es lo que saborea mucha gente, en una deliciosa taza de café. El café es conocido en Colombia. El café es famoso en todo el mundo.	¿Qué tomaba el abuelo? ¿Cómo se llamaba el pueblo? ¿A qué ayudaba el abuelo cuando era niño? ¿En qué se convierte la semilla cuando crece? ¿De qué color es el fruto que da el Cafeto? ¿Qué se hace primero: secar el café o tostarlo y molerlo? ¿Quién es el personaje principal de la historia? ¿Crees que esta historia podría suceder en la realidad?

Table B.4: Mathematics Test items

ID Scale	Comparing numbers	Ordering numbers	Equations	Additions	Subtractions
1	7 o 5 [7] 11 o 24 [24] 39 o 23 [39] 58 o 49 [58] 65 o 67 [67] 94 o 78 [94] 146 o 153 [153] 298 o 534 [534] 623 o 632 [632] 867 o 965 [965]	5 6 7 _ [8] 14 15 _ 17 [16] 20 _ 40 50 [30] _ 300 400 500 [200] 2 4 6 _ [8] 348 349 _ 351 [350] 28 _ 24 22 [26] 30 35 _ 45 [40] 550 540 530 _ [520] 3 8 _ 18 [13]	5 + 0 = _ [5] 6 + 1 = _ [7] 3 + 4 = _ [7] _ - 3 = 1 [4] 2+ _= 7 [5] 5+ _= 8 [3] _- 1 = 8 [9] _+ 0 = 8 [8] 8 - _ = 1 [7] _ + 4 = 9 [5]	2+2=(4)3+2=(5)4+2=(6) 1+5=(6)3+4=(7)7+1=(8) 6+2=(8)5+4=(9)4+5=(9) 7+2=(9)6+4=(10)5+5=(10) 8+2=(10)5+6=(11)6+6=(12) 3+9=(12)5+7=(12)8+6=(14) 10+3=(13)2+11=(13) 13+3=(16)6+10=(16) 10+10=(20)15+5=(20) 11+9=(20)	4-2=(2)8-1=(7)5-2=(3) 6-2=(4)8-2=(6)6-5=(1) 9-2=(7)9-4=(5)8-3=(5) 9-5=(4)7-4=(3)10-2=(8) 10-3=(7)10-4=(6)20-10=(10) 11-6=(5)11-7=(4)12-9=(3) 12-7=(5)12-6=(6)13-11=(2) 14-6=(8)16-3=(13)16-10=(6) 20-5=(15)20-4=(16) 20-9=(11)
2			15 + 15 = [30] 60- 20 = [40] 33 + 50 = [83] _+ 28 = 88 [60] 100 -_ = 10 [90] 29+ 61= [90] _+ 25 = 100 [75] _- 20 = 20 [40] 34 + _ = 100 [66] 32 + 39 = [71]		
3			48 + 52 = [100] 59 + 29 = [88] 28 + 74 = [102] _- 28 =4 [32] 5 x 2 = [10] 4 x 2= [8] _ x 10 = 30 [3] _- 50 = 100 [150] 5 x 4 = [20] 175 - _ = 150 [25]		

C Experiment Round II: Balance, Attrition, & Results

A second round of the experiment was run in 2019. Out of the 70 schools, half of them were assigned to each of the treatment and control groups. However, of the 70 schools initially in the experiment, three schools dropped-out for the second cohort (one treated and two controls).[27] Table C.1 presents balance and attrition for treated and control schools in the second cohort. Table C.2 exhibits the main results (pooled and by time horizon). We find that point estimates are large but not always significant at the standard levels. Columns 5 and shows that the intervention significantly improved knowledge of letter sounds and the composite literacy score, specially at end of first grade.

Table C.1: Balance and Attrition

		T	C	p-value	N
Panel A:	Avg. 2018 schools' characteristics				
	Rural school	0.29	0.24	0.26	67
	School size	292.85	350.97	0.17	67
	Class size	22.77	20.74	0.45	67
	Number of classrooms in Grade 1	1.44	1.36	0.77	67
	Literacy in Grade 3	0.64	0.64	0.95	67
	Math score in Grade 3	0.49	0.50	0.19	67
Panel B:	Avg. beg. G1 students' characteristics				
	Age	5.96	6.00	0.45	2253
	Female	0.46	0.46	0.99	2253
	No disability	0.97	0.96	0.39	2253
	Low socio-econ status	0.26	0.26	0.90	2253
Panel C:	Avg. end KG students' characteristics				
	Prob. response	0.67	0.67	0.60	2253
	Literacy	0.10	0.00	0.27	1505
	Motor	–0.00	0.00	0.96	1505
	Socio-affective	0.08	0.00	0.15	1505
Panel D:	Attrition				
	End Grade 1	0.12	0.16	0.16	2253
	Beg. Grade 2	0.28	0.35	0.32	1532

Note: Panel A shows the average pre-treatment characteristics of eligible students. Columns labeled 'T' show the average for students in schools randomized to treatment. Columns labeled 'C' show the average for students in schools randomized to control. We present these statistics by cohort.

[27] The dropped-out schools were schools that ended up not opening in 2019, i.e., the schools did not opted-out of the intervention.

Table C.2: Treatment Effects on Main Outcomes

	Knowledge of letter sounds	Reading of non-words	Reading passage	Reading comprehension	Literacy score
	(1)	(2)	(3)	(4)	(5)
Panel A: Pooled					
T	0.466***	0.067	0.090	0.163	0.206
	[0.128]	[0.109]	[0.116]	[0.124]	[0.137]
Panel B: By time horizon					
T x End G1	0.477***	0.129	0.121	0.187	0.249**
	[0.126]	[0.096]	[0.112]	[0.130]	[0.122]
T x Beg. G2	0.445***	–0.044	0.035	0.121	0.129
	[0.167]	[0.151]	[0.144]	[0.136]	[0.188]
Observations	3274	3274	3274	3274	3274
p-value Beg. G2 < End G1	0.400	0.036	0.166	0.231	0.152
Control mean at endline	13.94	15.88	32.04	3.266	65.12
Control s.d. at endline	12.99	12.22	24.74	2.607	44.70

Note: All models include cohort and strata fixed effects. Standard errors, shown in squared brackets, are clustered at the school level (the unit of randomization). The outcome scores were standardized with respect to eligible control students of respective year. * significant at 10%; ** significant at 5%; *** significant at 1%.