Álvarez Marinelli, Horacio; Berlinski, Samuel G.; Busso, Matias

Working Paper

Remedial education: Evidence from a sequence of experiments in Colombia

IDB Working Paper Series, No. IDB-WP-01067

Provided in Cooperation with:
Inter-American Development Bank (IDB), Washington, DC

Suggested Citation: Álvarez Marinelli, Horacio; Berlinski, Samuel G.; Busso, Matias (2019) : Remedial education: Evidence from a sequence of experiments in Colombia, IDB Working Paper Series, No. IDB-WP-01067, Inter-American Development Bank (IDB), Washington, DC, https://doi.org/10.18235/0002075

This Version is available at:
https://hdl.handle.net/10419/234669

IDB WORKING PAPER SERIES Nº IDB-WP-1067

Remedial Education:

Evidence from a Sequence of Experiments in Colombia

Horacio Álvarez Marinelli
Samuel Berlinski
Matías Busso

Inter-American Development Bank
Department of Research and Chief Economist

December 2019

Remedial Education:

Evidence from a Sequence of Experiments in Colombia

Horacio Álvarez Marinelli
Samuel Berlinski
Matías Busso

Inter-American Development Bank

December 2019

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

Álvarez Marinelli, Horacio.
Remedial education: evidence from a sequence of experiments in Colombia / Horacio Álvarez Marinelli, Samuel Berlinski, Matías Busso.
p. cm. — (IDB Working Paper Series ; 1067)
Includes bibliographic references.
1. Remedial teaching-Colombia. 2. Literacy-Colombia. 3. Academic achievement-Colombia. I. Berlinski, Samuel, 1970- II. Busso, Matías. III. Inter-American Development Bank. Department of Research and Chief Economist. IV. Inter-American Development Bank. Education Division. V. Title. VI. Series.
IDB-WP-1067

http://www.iadb.org

Abstract[1]

This paper assesses the effectiveness of an intervention aimed at improving the reading skills of struggling third-grade students in Colombia. In a series of randomized experiments, students participated in remedial tutorials conducted during school hours in small groups. Trained teachers used structured pedagogical materials that can be easily scaled up. Informed by the outcomes of each cohort, we fine-tuned the intervention tools for each subsequent cohort. We found positive and persistent impacts on literacy scores and positive spillovers on some mathematics scores. The effectiveness of the program grew over time, likely because of higher dosage and the fine-tuning of materials.

JEL classifications: C93, I21, J24, O15
Keywords: Remedial education, Literacy, Colombia, Sequential experimentation

[1] Álvarez Marinelli: Education Department, Inter-American Development Bank (horacioa@iadb.org). Berlinski: Research Department, Inter-American Development Bank (samuelb@iadb.org). Busso: Research Department, Inter-American Development Bank (mbusso@iadb.org). We thank Fundación Luker and the Secretaría de Educación Pública of Manizales for their support in implementing the intervention, in particular Santiago Isaza, María Camila Arango and Gloria de los Ríos. The pedagogical material was developed and revised over time by Alejandra Mielke, Eira Cotto, Angela Márquez and Mauricio Duque. We thank Jessica Gagete, Michele Giannola, Norbert Schady, and participants at various seminars for their useful comments. Anna Koh and Juanita Camacho provided excellent research assistance. The experiments in this paper have been registered in the AEA RCT Registry # AEARCTR-0005110.

1 Introduction

Literacy skills are essential for modern life. Literacy fosters the ability to learn other subjects (Zhang et al. [2014]). It matters for health (Sentell and Halpin [2006]) and political participation (Benavot [1996]), and it is highly valued in the labor market (Hanushek et al. [2015]). Yet UNESCO [2005] estimates that about 20 percent of the global adult population is illiterate. In developed countries, OECD [2016] finds that almost 20 percent of adults cannot process information from a simple text.[2] The problem is even more acute in developing countries. In Latin America, for instance, two-thirds of children do not achieve the minimum levels of literacy expected for their age (Busso et al. [2017]). The large number of children and adults struggling with reading demands attention and a remedy. In this paper we show how school-based, small-group tutorials can provide a remediation tool to close the literacy gap. We also demonstrate how policymakers can extract more benefits from their policies by using evidence to fine-tune current educational programs.

A growing literature in economics has recently focused on studying how changes in inputs affect literacy skills of school-age children. Three important insights about the effectiveness, and lack of effectiveness, of certain approaches have emerged so far. First, evidence suggests that providing access to books is usually an ineffective strategy for improving reading skills (Glewwe et al. [2009], Borkum et al. [2012] and Goux et al. [2017]) unless such increased access is accompanied by classroom strategies that encourage children to read (Abeberese et al. [2014]). Second, relatively small impacts on literacy scores result from interventions aimed at parents that provide them with literacy skills (Banerji et al. [2017]) or with information about their children's school performance (Barrera-Osorio et al. [2019]). By contrast, a third fundamental insight from the recent literature shows that certain teaching practices can have a profound, beneficial impact on how students acquire basic skills. Three characteristics seem to explain and underlie effective teaching: i) the use of structured materials for teaching reading (Machin and McNally [2008]), ii) the use of phonics-based methods for teaching reading (Machin et al. [2018] Hirata and e Oliveira [2019]), and iii) the use of content targeted at the right level of difficulty for the student when teaching reading and math (Muralidharan et al. [2019], Banerjee et al. [2017]). Our intervention includes these three teaching components.[3]

This paper presents experimental evidence on the impact of an intervention that offered remedial literacy tutorial sessions designed to help struggling readers in the third grade of primary school in Colombia. The tutorials consisted of 40-minute, structured sessions provided three times a week during the school day for up to 16 weeks. The sessions were

[2]Based on the Survey of Adult Skills, a product of the OECD Programme for the International Assessment of Adult Competencies (PIAAC), OECD [2016] found that "in almost all countries/economies, a sizeable proportion of adults (18.5 percent of adults, on average) has poor reading skills" (p.17). About 14.4 percent scored at level 1, and 4.5 percent scored below level 1 in the literacy test.

[3]Jacob [2017] evaluates the learning gains of students exposed to the Evidence-Based Literacy Instruction (EBLI), which provided teachers with several instructional strategies to improve reading accuracy, fluency and comprehension. The strategies were mostly based on phonics, but the curriculum was not structured. There was no significant difference in reading performance across the treatment and control classrooms.

conducted in small groups (six students maximum) and followed a simple structure. During each lesson tutors explained the objectives and activities, modeled the different exercises, and then used guided practice as well as student independent practice. The sessions used a curriculum that was designed and refined by international experts with support from a local team. The curriculum was based on a phonics approach. Lessons emphasized the ability to identify and manipulate units of oral language, the ability to recognize letter symbols and the sounds they represent, the ability to use combinations of letters that represent speech sounds, reading of words, reading fluency of sentences and paragraphs. It also worked on vocabulary and strategies for reading comprehension.

We evaluate the effectiveness of the intervention in more than 80 public schools of the Municipality of Manizales, Colombia, in three consecutive cohorts of third-grade students. Before randomization took place, we administered an initial literacy test to identify a total of 2,610 students who were struggling to read and who were thus deemed eligible to participate in the experiments. Most of these students lived in low-income households. Half of the schools were randomized into treatment and half into control groups. Tutors were hired, trained and randomly allocated to treatment schools.

We report six sets of results based on the experiment. First, we find that immediately after the experiment finished (at the end of the third grade) the overall literacy score of eligible students in treated schools improved by 0.286 standard deviations compared to the score of eligible students in control schools. The overall effect is explained by an increase in the ability of students to properly sound letters (0.356 standard deviations), an increase in the fluency of reading a paragraph (0.194 standard deviations), and a marginally statistically insignificant increase (0.073 standard deviations) in the reading of non-words. We find no effect on reading comprehension.

Second, the effects persist over time. We administered the tests at the beginning of fourth grade (about two months after classes had started) and at the end of fourth grade. By the end of fourth grade, there is a reduction of one-third on the estimated treatment magnitude for the overall literacy score. The effect of knowledge of letter sounds drops from 0.356 to 0.299, the effect on reading fluency is cut in half to 0.086, the effect on reading of non-words stays constant. However, we cannot reject the null hypothesis that in the three moments we measure our outcomes the effects are the same.

Third, we administered a standardized math test. We find that treated children performed better in addition problems both immediately after the literacy treatment finished and during the exams administered in fourth grade. The gains range between 0.077 and 0.108 of a standard deviation. We also find positive but not statistically significant effects on subtractions (with treatment effects between 0.030 and 0.098). Despite these gains in both literacy and math, treated children were equally likely to repeat third grade.

Fourth, these treatment effects are homogeneous in key respects. Following Firpo [2007] we estimate fairly constant quantile treatment effects for most outcomes of interest. In addition, we explore heterogeneity of treatment effects by tutorial characteristics. We find no

significant heterogeneity across students attending smaller versus large tutorials, or having worse versus better peers, or being in a more homogeneous versus a more heterogeneous group. The only dimension that seems relevant is whether the tutor has previous experience with the program. Students of new tutors seem to gain relatively more in properly sounding letters (a relatively simple subtask) while students of more experienced tutors do better in reading of non-words and reading fluency (arguably more complex subtasks).

Fifth, the effectiveness of the intervention increased over time. The median effect estimated over all outcomes and grades increases from 0.015 of a standard deviation in cohort 1, to 0.137 in cohort 2, to 0.204 in cohort 3. The gains by the end of the third grade on the aggregate literacy score goes from 0.120 in cohort 1, to 0.225 and 0.609 in cohorts 2 and 3. A similar pattern is observed for each of the subtasks. These results can be explained by deliberate refining of the program. Feedback from each cohort was used to improve the intervention effectiveness in the next wave of the intervention.

Finally, we present several back-of-the-envelope calculations to quantify the changes that could explain the increased effectiveness of the intervention over time. The analysis suggests that the increase in dosage (more sessions and higher attendance rates) plays an important role. Other factors such as the targeting of the intervention, the composition of the tutorial groups, and the increased experience of the tutors seem to be of less importance. Some of the difference might be attributed to the fine-tuning of the material, but this is difficult to quantify. We speculate that this is likely the main driver of the differences in the effectiveness of the intervention between the second and third cohorts.

Our results are directly relevant for policy debates regarding the timing and effectiveness of human capital interventions. At the core of the definition of developmental milestones is the idea that stages of development occur during predictable time periods. Child development specialists have long studied whether there are critical and sensitive periods for physical and skills development. Cunha and Heckman [2007], among others, have argued that there might be a sensitive age range in which achieving a certain trait or skill requires fewer resources, and alternatively, that the absence of some experience in a certain age range may have permanent developmental consequences. A classic example of a sensitive age range refers to the acquisition of vision. Though educational experts have suggested that reading is best acquired in the very early elementary school years, our experiment shows that an easily scalable literacy remediation program can have significant impacts even by the end of third grade. Our paper is close to the approach of Banerjee et al. [2007] that analyzed randomized, clustered evaluation of an intervention targeted to low achieving students. The intervention described in their paper is much more intensive than ours. They report results of a year-long program that provided struggling students (as determined by the schools) with two hours a day (i.e., half day) of tutoring support in literacy and math in the third and fourth grade by an external tutor ("Bhalsaki"). The study finds that on average students in treated schools improved literacy test scores with the effect decaying over time. A limitation of this study is that there is no information on ex ante eligibility for children in the control units.

Our results are also relevant for policy discussions regarding the process used in the design

of social policies. There are many small-scale studies in developed countries that look at reading remediation early in elementary school. A meta-analysis (Slavin et al. [2009]) of this work identifies teacher development and phonological awareness as successful practices. It is an open question whether remediation will work at scale, particularly in developing countries with less qualified teachers (Kerwin and Thornton [2019]). Our paper shows that it does. Finally, economists are increasingly interacting with policymakers in the selection and design of policies, and, as a result, the economist mindset can affect the learning that happens in the course of the experiment itself (Duflo [2017]). Our paper is an example of such learning. By using the results observed in each cohort and working with feedback from experts in the field over the course of the experiment, we were able to increase the intervention's effectiveness over time.

The rest of the paper is organized as follows. Section 2 describes the intervention and the setting in which it took place. Section 3 presents the experiment and the data. Section 4 presents the main results of the paper and discusses implications. Section 5 presents results that explain the increased effectiveness of the intervention over time. Section 6 provides calculations of the cost-effectiveness of the intervention. Section 7 concludes.

2 Intervention

2.1 Setting

The sequence of experiments took place among third grade students of public elementary schools in the Municipality of Manizales in Colombia during three consecutive years (2015-2017). About 97 percent of the children participating in the study can be considered as disadvantaged.[4] Manizales is a mid-size city. Approximately 13.8 percent of residents had incomes below the poverty line, and 6.9 percent of the municipality's residents lived in rural areas. More than 18,000 children were enrolled in the first five grades of the public elementary school system.[5] The Municipality scored slightly above the national mean among third-graders in the 2016 national standardized language achievement tests. However, almost 45 percent of students scored at or below the "minimal knowledge" threshold.

The Secretary of Education of Manizales, in partnership with a local NGO (Fundacion Luker), implemented a series of interventions aimed at improving the poor results on standardized tests. A first step in this direction was to create a remedial program to improve reading fluency among struggling third-grade students. Fluency is a good indicator of reading proficiency because it is associated with comprehension in novice readers (Fuchs et al. [2001]). Good et al. [2001] report that 96 percent of children who met the third-grade oral reading fluency benchmark goal, also met or exceeded expectations in a high-stakes,

[4]In our sample, 97 percent of students fall in levels zero to three of the social stratification classification scale used in Colombia to target social programs.

[5]Schooling in Colombia is compulsory from kindergarten to Grade 9. Both public and private schools operate in Colombia, and about 78 percent of school-aged children in the Municipality of Manizales attend public schools. Most children in our sample attended the school closest to their home. Schools operate in either six- or eight-hour schedules for 165 days a year.

statewide assessment. Furthermore, the importance of fluency and basic reading skills goes beyond elementary school. Shaywitz and Shaywitz [1996] found that 74 percent of children who were poor readers at the end of third grade were likely to still be poor readers by the time they reached the end of ninth grade.

2.2 Reading Remediation in Small Tutorial Groups

There is a strong consensus from research on reading instruction (see, for example, Foorman and Torgesen [2001] and NAEP [2000]) about the necessary skills that children should develop in the early years of school: phonemic awareness (i.e., the ability to identify and manipulate units of oral language), decoding skills (i.e., the ability to recognize letter symbols and the sounds they represent), fluency in word recognition (i.e., the ability to read with speed, accuracy, and proper expression), text processing, construction of meaning, vocabulary, spelling, and writing skills.

Ehri [2005] describes the process of reading as one in which connections are made that link the spelling of written words to their pronunciation and meaning in memory. In an initial phase, children learn the names or sounds of letters of the alphabet and use them to learn how to read words. Children use these tools to learn new words that they can, through repeated use, then recognize as a unit by sight. To construct meaning from texts, students need foundational skills, including phonological awareness, decoding and fluency. Good et al. [2001] propose a timeline for the development of these skills: phonological awareness during kindergarten, decoding and acquiring the alphabetical principle in first grade, gaining accuracy and fluency when reading in second and third grades. Students lacking these skills by third grade will not be able to read fluently. Longitudinal studies show that students with poor reading skills in earlier grades do not catch up with their peers who are good readers. In fact, the gap in the developmental reading trajectories of poor readers versus more proficient ones keeps expanding over time (Good et al. [1998], Stanovich [1986]).

Children at risk of reading failure acquire reading skills more slowly than other children. According to Foorman and Torgesen [2001], instruction for these children must be phonemically more explicit (i.e., use systematic instruction to build phonemic awareness), more intensive, and more cognitively supportive (i.e., provide carefully scaffolded instruction). To achieve these objectives, we produced a highly structured intervention delivered in small-group tutorials to achieve the required intensity.

At the beginning of each lesson the tutor explained to the students the learning outcomes, objectives, and activities for each session. The tutor modeled the different exercises, and then used guided practice (i.e., tutor practices the target ability with students) and independent practice (i.e., students practice the target ability on their own and/or in pairs) to foster learning among students. Both tutors and students received a workbook as part of the intervention. Scaffolded lessons emphasized phonological awareness, decoding, alphabetic principles (i.e., the ability to use diagraphs, which are letter combinations that represent speech sounds in a predictable and systematic way), vocabulary, reading fluency strategies, and comprehension strategies. Each 40-minute session was designed to dedicate 20 minutes

to reading fluency-related exercises, 10 minutes to vocabulary building, and the other 10 to reading comprehension strategies.

The intervention was implemented during the second half of each academic year (starting right after the June mid-year break). Sessions were delivered three times a week. During the first cohort, the intervention lasted for 36 tutorial sessions (12 weeks in total). In the last two cohorts of the experiment this was extended to 48 sessions (16 weeks in total).

The most practical method for increasing instructional intensity for a small number of at-risk students is to provide small-group instruction. Meta-analyses in education (see,Foorman and Torgesen [2001], Inns et al. [2019]) consistently find positive impacts of small-scale, well-designed interventions in which students are taught in groups of two to six students. Although the evidence is not yet overwhelming (see Elbaum et al. [2000]), an interesting finding that has been emerging from these analyses is that one-to-one interventions in reading are not necessarily more effective than small-group interventions. There is also evidence in education (see, Elbaum et al. [2000]) that many successful interventions can be delivered by trained individuals rather than reading specialists.

Struggling readers were taken out of the classrooms during regular school hours. Tutors led tutorials for a small group of students (no more than six students) in a designated school space. Fifteen tutors were hired each year of the intervention.[6] These were trained primary school teachers, psychologists and audiologists with some teaching experience. The average tutor was 29 years old, and 97 percent of them were women. Each tutor oversaw an average of five tutorial groups each year. At the beginning of each year of the intervention, tutors received an eight-hour training session. Once the program was under way, the tutors participated in regular meetings for coaching and feedback, and they were observed during two on-site supervised sessions (by their trainers).

3 Research Design

3.1 Measures

We measure language development using the Early Grade Reading Assessment (EGRA). Designed by RTI-International [2009] under the auspices of 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 and Gove [2015]). EGRA is a research-based collection of individual subtasks that measure some of the foundational skills needed for reading acquisition in alphabetic languages (Dubeck and Gove [2015]; p. 317). 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.[7]

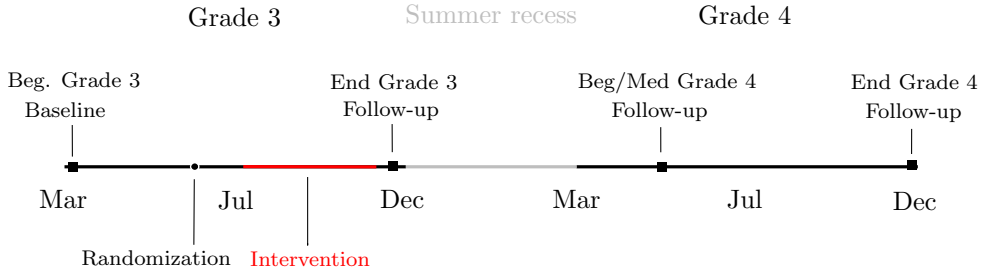
[6]During the three years in which the intervention was implemented, the program hired a total of 33 tutors. One third of them participated in multiple rounds.

[7]For most subtasks, the items within it are, a priori, of equal difficulty.

We collected information on the following EGRA subtasks: knowledge of letter sounds, reading of non-words, fluency of oral reading, and reading comprehension. We also used the Early Grade Math Assessment (EGMA) to assess early grade mathematical competence. We focus on subtasks that measure addition and subtraction of one- and two-digit numbers. Both tests were administered orally by trained enumerators, one-on-one with a child, using a tablet. The application of the tests takes less than 20 minutes per student. The tests were applied to the universe of children in public schools. In the Data Appendix we report the test-items administered at every point in time for each subtask, as well as their psychometric properties.

Figure 1 describes the timeline for data collection and other activities related to the experiment for each of the three cohorts of students. At the beginning of the academic year we collected information about students in Grade 3. This is our baseline. In addition, to measure the impact of the intervention, we administered the instruments to the same population of children at the end of Grade 3 and at the beginning/middle and end of Grade 4.

Figure 1: Timeline (for Each Cohort)



Note: The figure shows the timeline of intervention and data collection for the three experiments implemented in 2015, 2016 and 2017.

3.2 Sample

We used the information collected at the beginning of the school year from the universe of schools in the municipality to determine the number of eligible children in each school. We sorted schools based on how many children were eligible for treatment. In the second and third cohorts we eliminated schools with too few (less than two) or too many (more than 35) eligible students.[8] We created blocks of two and within these strata randomized schools to treatment and control status.

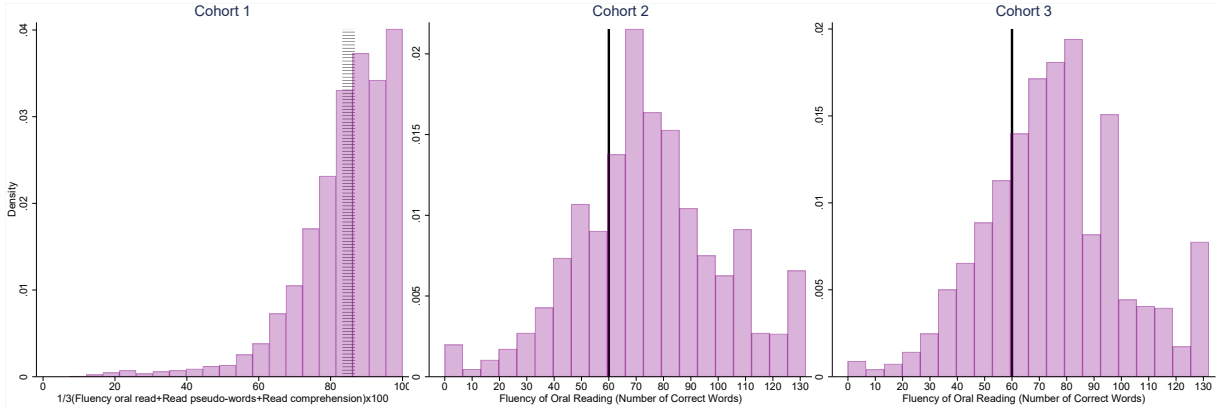
In the case of the first cohort, students were eligible if they scored in the bottom 25 percent of an equally weighted composite index of the following EGRA subtasks: reading of non-words, fluency of oral reading, and reading comprehension. We changed the eligibility criteria during the second and third cohorts. We established that children would be eligible for treatment if they correctly read fewer than 60 of the 132 words in a paragraph in the EGRA fluency of oral reading subtask. Figure 2 depicts the distribution of baseline scores

[8]During the first cohort we found that that groups that were too small or too large added too much logistical complexity to the intervention.

and the threshold for eligibility in the universe of children.

The Data Appendix presents information on the sample sizes and the response rates of each cohort. We started with a universe of 94 schools in 2015. We reduced the experimental sample to 84 schools largely due to logistical considerations. In the two subsequent years the experimental sample had a total of 80 schools.

Figure 2: Eligibility Criteria



Note: The eligibility criteria changed between the first and the second cohort. For cohort 1 eligible students were defined as those that performed in the bottom 25 percent of the study population distribution of a composite literacy score. For cohorts 2 and 3 eligible students were defined as those that read fewer than 60 words in the fluency of oral reading EGRA subtask. The fuzzy vertical line in the left figure represents the fact that in the case of cohort 1, tutorials were completed to maximum capacity (of size 6) with students immediately next to the 25th percentile threshold. In cohorts 2 and 3 the eligibility criteria were strict (represented by a solid line).

3.3 Randomization

We randomly assigned schools to treatment and control status in each of the three cohorts of the intervention (i.e., treatment and control schools might differ each year). Eligible children in treatment schools participated in the remedial reading program while those in the control schools carried on with their usual classroom learning experiences.

Tutors were randomly assigned to schools. In each school there was only one tutor. In the case of the first cohort, when there were more than six eligible children, tutors and schools organized the compositions of the tutorials. In schools with fewer than six students for a given tutorial, those close enough to the cut-off participated to fill the session with up to the maximum of six students. This resulted in tutorials with one more student, on average, in cohort 1. We modified the tutorial assignments in the second and third cohorts: in schools with more than six eligible students, students were assigned randomly to equally sized tutorials. Students above the cut-off were not offered treatment.

3.4 Experimental Validity

We assess the experimental validity of our research design by estimating the difference (θ) in pre-treatment characteristics, attrition, and treatment compliance between eligible students enrolled in treated schools and eligible students attending a control school. For each cohort we estimate θ using an OLS model of the form:

$$W_{is} = \theta T_s + \mu_{\text{strata}} + \epsilon_{is} \tag{1}$$

where W_{is} is a variable of interest, T_s an indicator variable equal to one if the student i is enrolled in school s that was randomized into treatment, μ_{strata}, and is a strata fixed effect. Standard errors are clustered at the school level, which served as the unit of randomization.

Table 1: Balance, Attrition, and Compliance

			Cohort 1			Cohort 2			Cohort 3		
		All	T	C	p-value	T	C	p-value	T	C	p-value
Panel A:	Avg. pre-treatment characteristics										
	Age	8.61	8.82	8.83	0.71	8.49	8.37	0.09	8.40	8.47	0.31
	Proportion female	0.49	0.50	0.48	0.74	0.47	0.50	0.42	0.50	0.52	0.76
	Prop. low Socio-ec. status	0.35	0.36	0.37	0.90	0.35	0.33	0.64	0.31	0.35	0.22
	Scores:										
	Fluency of oral reading	46.17	49.84	50.73	0.14	42.17	42.12	0.98	44.17	43.42	0.51
	Knowledge of letter sounds	15.22	14.82	13.64	0.20	12.41	13.46	0.18	19.51	19.00	0.69
	Reading of non-words	26.90	28.83	28.33	0.54	24.22	24.88	0.41	26.62	26.83	0.84
	Reading comprehension	3.36	3.07	3.10	0.55	2.10	1.99	0.20	5.33	5.05	0.26
	Literacy score (avg. of sub-scores)	0.51	0.59	0.59	0.90	0.39	0.39	0.37	0.49	0.48	0.55
	Additions	11.49	12.18	12.23	0.60	10.15	9.97	0.57	11.78	11.89	0.67
	Subtractions	8.93	9.51	9.21	0.40	8.12	7.74	0.26	9.44	9.16	0.35
	Math score (avg. of sub-scores)	0.40	0.42	0.42	0.95	0.35	0.34	0.31	0.41	0.41	0.87
Panel B:	Attrition										
	End of Grade 3	0.08	0.06	0.07	0.19	0.07	0.09	0.27	0.09	0.09	0.95
	Beg./Mid. Grade 4	0.32	0.54	0.60	0.25	0.12	0.16	0.08	0.10	0.12	0.40
	End Grade 4	0.17	0.21	0.17	0.09	0.14	0.12	0.50	0.14	0.18	0.20
Panel C:	Compliance										
	Ever attended a tutorial	0.47	0.92	0.00	0.00	0.94	0.00	0.00	0.98	0.00	0.00
	Tutorial attendance (percent)	0.41	0.73	0.00	0.00	0.90	0.00	0.00	0.90	0.00	0.00
	Tutorial attendance (sessions)	17.69	26.20	0.00	0.00	43.03	0.00	0.00	42.99	0.00	0.00
	Average tutorial size	5.34	5.92			4.82			4.99		

Note: Panel A shows the average pre-treatment characteristics of eligible students. Scores for subtasks are expressed in number of correct responses (fluency of oral reading, reading of non-words, reading comprehension, additions, subtractions). Literacy and math scores are expressed as average proportion of correct answers in each sub-task. Panel B shows the attrition rates at different time horizons of the eligible students observed at baseline. Panel C shows which eligible students attended the tutorials and the intensity of attendance. Column labeled 'All' shows the average across the three cohorts. 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. Columns labeled 'p-val' show the p-value of a test of $H_0 : \theta = 0$ (see equation (1)). We present these statistics by cohort.

Panel A of Table 1 shows students' observable characteristics at the beginning of the school year, before each round of the experiment for each cohort. The students in the experiment were, on average, 8.6 years old. Half of them were female, and 35 percent were from low-socioeconomic status households. Demographic characteristics, as well as reading and math scores at baseline, were not statistically different for students in treatment and control schools in any of the three cohorts.

We collected information about eligible students at different points in time. For this reason, it is important to assess the level of differential attrition between treated and control schools. Most of the attrition observed in our sample was caused by students not being in school on the day the exam was administered. (Very few students dropped out of schools.) Panel B of Table 1 shows the probability that a student deemed eligible to receive treatment at baseline failed to take an exam on each subsequent date. All in all, there is no evidence of differential attrition between treatment and control schools. However, it is important to note two things. First, the attrition rate for the first cohort in the first measure of Grade 4 is 54-60 percent. This was due to a logistical problem with the data collection that prevented administering the test in several schools. Second, we do reject at the 10 percent level the null hypothesis of equality of attrition for two of the nine tests. In those cases, the differences in the rates are close to 4 percent higher in the control group. In the Data Appendix we show that the characteristics of students that attrited from the sample are essentially the same as those of students who did not.

Panel C of Table 1 shows the participation of students in the tutorial groups for both treatment arms. For all three cohorts of students in the control schools, the attendance rates were zero, for the simple reason that the tutorials were not offered in those institutions. Attendance in treated schools was high. On average, students in the first cohort attended 73 percent of the offered tutorials. Students in the second and third cohorts of the experiment had an attendance rate of 90 percent.

4 The Causal Impact of Remediation in Small-Group Tutorials

In this section we report the estimates of the intention-to-treat effects for the eligible population enrolled in schools that were randomized into treatment.

4.1 Empirical Strategy

We have multiple measures of the same outcomes at the beginning/end of third/fourth grade for three cohorts of children. We stack this information and then estimate the following model:

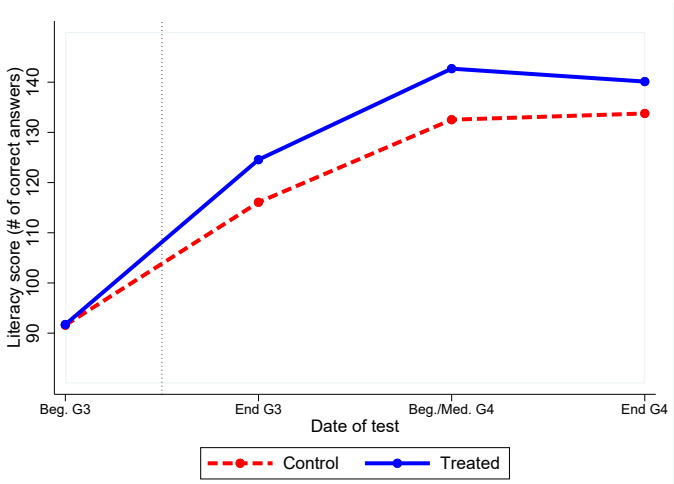
$$Y_{isch} = \alpha + \sum_{h=1}^{3}(\theta_h \times P_h \times T_{sc}) + \mu_c + \gamma_h + \epsilon_{isch} \tag{2}$$

where Y_{isch} is an outcome for student i who attends school s, and belongs to experimental cohort c. This outcome is measured at time horizons h – that is, at the end of the third grade ($h = 1$), at the beginning/middle of fourth grade ($h = 2$), and at the end of fourth grade ($h = 3$).[9] μ_c is a strata fixed effect defined for each cohort c at the time of school randomization into treatment and γ_h are time horizons fixed effects (with $h = 1$ excluded). T_{sc} an indicator variable equal to one if the student was enrolled in third grade at a school s randomized into treatment in cohort c and P_h is an indicator which takes value of one for each time horizon h. Thus, the parameters of interest are θ_h, which measure the intention-to-treat effect at $h = 1, 2$ and 3. Standard errors are clustered at the school level, the unit of assignment to treatment.

4.2 Main Results

We plot in Figure 3 the raw count of correct answers over all literacy subtasks, aggregating the information for the three cohorts. At the beginning of third grade, on average, students in treated and control schools correctly answered 91 items. The control group correctly answered 116 items at the end of grade three, and 133 items by the end of grade four. By contrast, students in treated schools correctly answered 124 (end of Grade 3) and 140 items (end of Grade 4). The figure suggests that the treatment group experienced positive gains from this intervention, and that the gains persisted over time.[10]

Figure 3: Effect of the Intervention on Literacy Scores



Note: The solid line shows the number of correct answers by eligible students in schools randomized to treatment. The dashed line shows the number of correct answers by students in schools randomized to control. "Beg G3" refers to the measure taken in March (baseline) of Grade 3, "End G3" refers to the measure taken by the end of Grade 3, after treatment. "Beg/Med G4" refers to the first measure taken in Grade 4, "End G4" refers to the measure taken by the end of Grade 4. The vertical dotted line marks the approximate time of treatment.

[9] Some students repeat, and therefore they are observed twice in the third grade. Thus, the model also includes year dummies.

[10] The growth rate is much higher between the beginning and end of the third grade than in the fourth grade because subtasks get relatively easier once a minimum of reading fluency is achieved.

Table 2 presents the main results of the paper. We show the intention-to-treat effect on each of the subtasks: knowledge of letter sounds, reading of non-words, fluency of oral reading, and reading comprehension. We also include a literacy score, which is the sum of correct answers across all subtasks. All outcomes are standardized by the mean and standard deviation observed in the control group of each cohort at the corresponding point of measurement.

We start by estimating the impact on knowledge of letter sounds. Children in our control group properly sounded an average of 15 letters at the beginning of the experiment. At the end of the third grade, we estimate that the causal impact of the program is 0.356 standard deviations (or four letter sounds).[11] We look next at non-word reading, which measures the ability to decode individual non-words that follow a common orthographic structure. At baseline the control group children read correctly an average of 27 non-words. The treatment effect is 0.073 standard deviations which translates into a gain of less than an extra non-word. Column 3 estimates the impact on oral reading fluency, which measures the ability to read a grade-level text. In the control group, children correctly read on average 46 words. We found that treated children's reading scores were 0.194 standard deviations higher (representing a gain of almost three words) than those of the control group. Column 4 shows the results for reading comprehension, which measures the ability to answer explicit, inferential, and look-back questions about the grade-level text student had just read for the fluency of oral reading subtask.

We do not find any impact on reading comprehension. Broadly speaking children that become successful readers bring to schools two sets of skills (see, Whitehurst and Lonigan [1998] and Foorman and Torgesen [2001]). One of them involves the ability to manipulate letters, sounds and phonemes. The other includes vocabulary and conceptual knowledge. Both are key to ultimately reading with meaning. Our relatively short intervention is focused on improving reading fluency. However, if children from disadvantaged backgrounds are also impoverished in the quality of verbal interactions with adults (Hart and Risley [1995]), which affects vocabulary and conceptual knowledge, improving reading comprehension may require a longer intervention that places appropriate emphasis on these aspects of literacy development.

We summarize the effects on reading in an overall literacy score –the proportion of correct answers in all subtasks- which shows an impact of the intervention by the end of third grade of 0.286 standard deviations. We take this effect to be quite large considering that the gain during third grade of the average student in the control group is 0.400 standard deviations.

4.3 Medium-Run Results

If gained skills are not used or reinforced, the impact of programs tends to decline over time; thus, fade-out is common in early childhood and education interventions (e.g., Currie and Thomas [1995], Deming [2009], and Chetty et al. [2011]). Because reading is a skill that

[11] Note that the number of observations in "Knowledge of letter sounds" is smaller than that of the other outcomes. This is because we did not test letter sounds in grade four for the first cohort.

Table 2: Treatment Effects on Main Outcomes

	Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading comprehension	Literacy score
	(1)	(2)	(3)	(4)	(5)
Treatment x End of Grade 3	0.356***	0.073	0.194***	0.005	0.286***
	[0.059]	[0.049]	[0.050]	[0.035]	[0.056]
Treatment x Beginning of Grade 4	0.323***	0.104*	0.186**	0.116**	0.271***
	[0.070]	[0.054]	[0.076]	[0.051]	[0.076]
Treatment x End of Grade 4	0.299***	0.086**	0.086	0.042	0.164***
	[0.056]	[0.041]	[0.052]	[0.042]	[0.052]
Observations	4949	6362	6362	6362	6362
p-value of equal coeffs.	0.651	0.881	0.136	0.137	0.060
Mean	15.02	26.94	46.32	3.303	0.507
S.D.	11.36	10.04	13.81	1.877	0.142

Note: 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 row labeled 'p-value of equal coeffs' shows the p-value of a test $H_0 : \theta_1 = \theta_2 = \theta_3$ and the rows labeled 'Mean' and 'S.D.' show the average and standard deviation of the outcomes of students in the control group at baseline. All models include cohort, year, and strata 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%.

children may easily use outside the classroom, it is a priori less clear whether the gains of the intervention will fade out during fourth grade. Even though the point estimates fall slightly, it is hard to reject the null hypothesis that the magnitude of the effects are similar in the three periods in which we measure the impacts. There is clearly no fade-out in letter-sounds knowledge and reading of non-words. The magnitude of the effect on fluency of oral reading is smaller at the end of fourth grade, but we cannot reject the null hypothesis that the impact is the same in the third grade. The positive effect on reading comprehension toward the beginning of fourth grade is statistically different than the smaller impacts at the end of third grade and fourth grade. It is hard to speculate why this happens.

We summarize the impact on literacy using an index that adds the number of correct responses in each subtask. Overall, we find a gain of 0.286, 0.271 and 0.164 of a standard deviation in each of the three impact measurement episodes. All of them are statistically significant at the 1 percent level. These results show that a small-group tutorial designed to help struggling readers improved reading skills and that its effects persisted.

4.4 Robustness

In Appendix Table A.2 we show that our main results are robust to several changes in the model specification. First, we add controls to equation 2. Neither including the corresponding baseline test scores (panel A) nor including individual and school controls (panel B) affects the results. This is not surprising given that we have shown that these variables are balanced before treatment. Further, we condition on school fixed effects (panel C) by exploiting the fact that 50 schools changed treatment status during the three rounds of the

experiment.[12]Reassuringly, the intention-to-treat estimates remain unchanged.

4.5 Effects on Other Outcomes

The tutorial required students to be taken out of the classroom which could have negatively affected their classroom learning by receiving fewer hours of instruction from their main teacher. On the other hand, improved literacy skills may have had positive impacts on other subjects by potentially enhancing students' ability to follow instructional materials, and perhaps, indirectly, through improved self-esteem. Machin and McNally [2008] and Machin et al. [2018], for example, find spillovers to mathematics from interventions that successfully change reading skills.

Table 3: Treatment Effects on other Education Outcomes

	Additions	Subtractions	Math score	Repeat Grade 3
	(1)	(2)	(3)	(4)
Treatment x End of Grade 3	0.077	0.030	0.061	-0.005
	[0.053]	[0.054]	[0.054]	[0.013]
Treatment x Beginning of Grade 4	0.108*	0.098	0.117*	
	[0.055]	[0.065]	[0.062]	
Treatment x End of Grade 4	0.092**	0.064	0.088*	
	[0.045]	[0.048]	[0.046]	
Observations	6362	6362	6362	2391
p-value of equal coeffs.	0.878	0.708	0.723	
Mean	11.49	8.778	0.393	0.129
S.D.	4.803	4.381	0.157	0.335

Note: Each column shows the estimates of the coefficients β_h of equation (3), that is, the estimated dose response at different time horizons for each outcome of interest. Dosage is measured by the number of days in which students attended the tutorial. The actual attendance was instrumented with the randomized treatment indicator variable. The row labeled 'p-value of equal coeffs.' shows the p-value of a test $H_0 : \beta_1 = \beta_2 = \beta_3$. The average and standard deviation of the outcomes of students in the control group at baseline can be seen in Table 2. All models include cohort, year, and strata 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%.

Table 3 investigates the effect of the intervention on other outcomes not directly targeted by the material used in the reading tutorials. We find positive and statistically significant effects on students' ability to solve simple one-digit addition problems. The effect on the subtraction subtask is similar in magnitude but not statistically significant. Overall the intervention had a positive effect on math scores. The magnitude is between one-quarter and one-third of the effect on literacy. Despite the learning gains and the fact that around 13 percent in the control group repeat the grade, we do not find that the intervention affects the probability of repeating the grade.

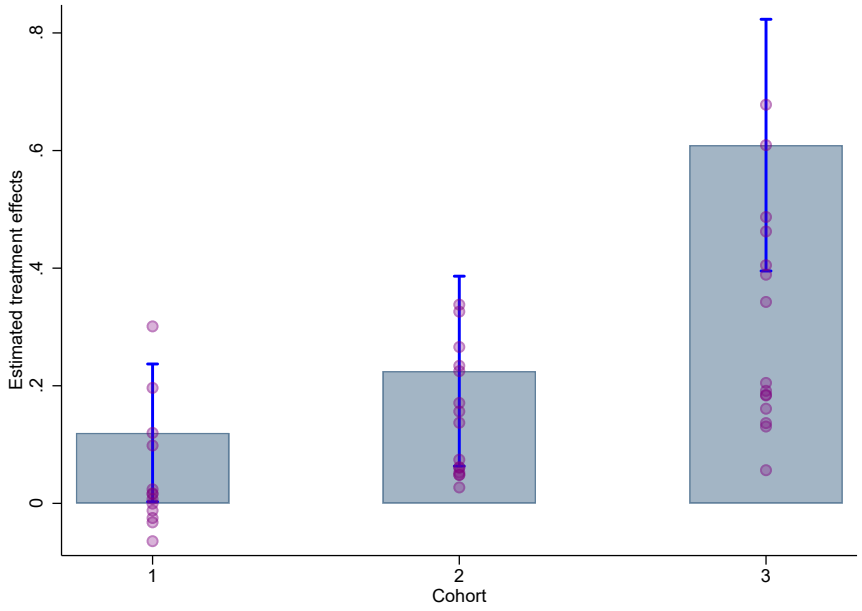
[12]In results not shown, and available from the authors upon request, we find that the intervention did not affect the probability of changing schools.

5 Sequential Experimental Results

The results summarized in the previous section are constructed from a sequence of experiments. The feedback from each cohort was used to improve learning in the next wave of the intervention. It is therefore instructive to examine how the treatment effects vary across cohorts, and to study the potential channels that can explain the improvements we document in this section.

We estimate the treatment effects for each cohort, outcome, and time horizon combination. A total of 43 parameters are reported in Appendix Table A.1. Figure 4 summarizes this information using a box-plot where the size of the box measures the inter-quartile range of the estimates, and the line inside the box shows the median estimate.[13] There is an upward trend over time in the treatment effect of the intervention. The median effect increases from 0.015 of a standard deviation in cohort 1, to 0.137 in cohort 2, to 0.204 in cohort 3. The gains by the end of the third grade on the aggregate literacy score grow from 0.120 in cohort 1, to 0.225 in cohort 2, and 0.609 in cohort 3. A similar pattern emerges for each subtask. It is reassuring that reading of non-words, the only subtask that had the same test items across all cohorts, increased from 0.024, to 0.053, to 0.183 from cohorts 1 to 3.

Figure 4: Treatment Effects by Cohort



Note: Each bar shows the estimated treatment effect for the aggregate literacy score for each cohort. Each bar presents the corresponding 95 percent confidence interval. In addition, circles present the estimated treatment effects for each literacy subtask, estimated at each time horizon for each cohort. See Appendix Table A.1 for the individual estimated treatment effects.

Our analysis of results from the first cohort showed positive treatment effects on knowledge of letter sounds only (see Appendix Table A.1). These results were to some extent disap-

[13] Results are similar when the estimates are weighted by the inverse of the standard error.

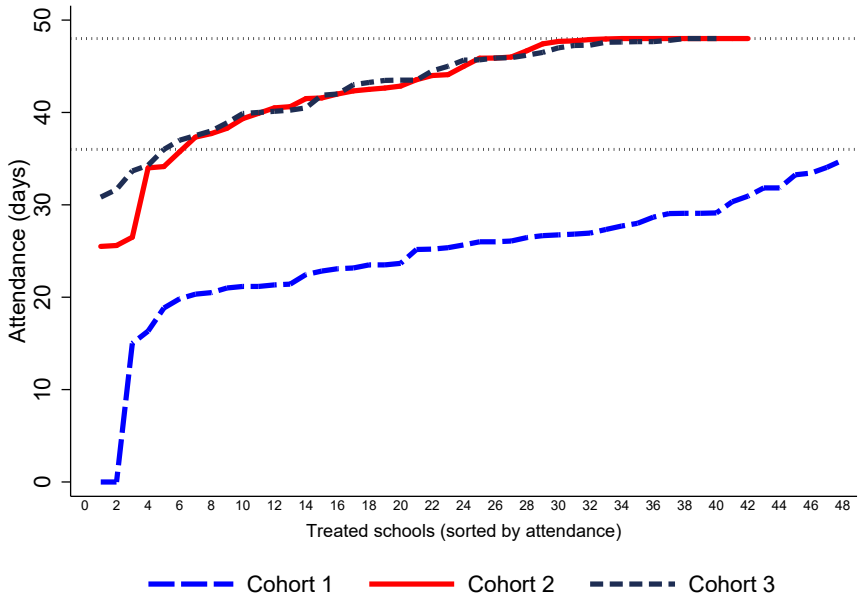
pointing because we did not manage to improve our main target outcome, reading fluency. As a consequence, the research and the implementation teams identified several areas where we could improve the intervention for the following rounds. The focus of these changes was to increase intensity and to improve the cognitive support of the intervention. In order to increase intensity we introduced make-up sessions, focused our targeting on those that exhibit the poorest results in reading fluency, and reduced tutorial size. To improve cognitive support we increased the number of sessions and we review the pedagogical material by replacing some of the vocabulary development tasks in favor of exercises that promoted reading fluency. With an eye at improving the scaffolding of the intervention we reorganized the readings in some sessions and adjusted the difficulty of the texts.

In other words, we introduced four changes after the first cohort: we increased the dosage, we modified the targeting, we modified the assignment of students to tutorial groups, and we fine-tuned the material. Additionally, in the wake of the experiment with the first cohort, tutors for cohorts 2 and 3 had previous experience in delivering the intervention. Next, we analyze how these five factors may have contributed to the increased impact of the intervention over time.

5.1 Dosage

Figure 5 presents the average number of days that students in each cohort attended a tutorial session (sorted from low to high attendance). We increased the dosage of the intervention by increasing the number of tutorial sessions from 36 to 48 (marked as dotted lines in the figure). This generated a clear upward shift in the number of attended tutorials between the first cohort and the subsequent cohorts. We also introduced make-up sessions to provide better coverage of the course material. These make-up sessions, administered by the same tutors, allowed students who missed a tutorial class to cover the relevant material so as not to fall behind with respect to their small-group tutorial peers. Even though we observe variation in the attendance rates by schools in all three cohorts, this make-up option led to perfect attendance at more schools for the second and third cohorts.

Figure 5: Attendance at Tutorials



Note: Each line shows the average number of days attended by students in each cohort. In each cohort, we sorted the schools from lowest to highest attendance. The line for cohort 1 spans more schools because the sample of schools in the experiment in cohort 1 was larger than that in cohorts 2 and 3 (see Section 2 for more details). Horizontal lines show the total number of tutorials offered: 36 for students in cohort 1 and 48 for students in cohorts 2 and 3.

To measure the contribution of increased attendance to the different treatment effects by cohort, we estimate dose-response effects using the following model:

$$Y_{isch} = \alpha + \sum_{h=1}^{3}(\beta_h \times P_h \times D_{isc}) + \mu_c + \gamma_h + \epsilon_{isch} \tag{3}$$

where D_{isc} is the number of tutorials attended by student i in school s from cohort c, P_h is an indicator variable equal to one when the outcome is measured at time horizon h, and β_h captures the dose-response effect at time horizon h (with $h = 1, 2, 3$). Similar to equation (2) we include strata-cohort and time-horizon fixed effects. Because attendance at tutorials might not be orthogonal to ϵ_{isch}, we instrument it with the randomized treatment variable T_{sc} (interacted with the time-horizon indicator variables).

Table 4 shows that there is a positive dosage effect. At the end of third grade, students who attended one additional session performed 0.008 of a standard deviation better in the literacy test than students in the control group.[14] These results decay slightly in fourth grade, but we cannot reject the null hypothesis of equality of coefficients. Students in cohorts 2 and 3 attended on average 17 more sessions than students in the first cohort (see Panel C of Table 1). This would translate into a gain of $0.008 \times 17 = 0.136$ which is similar to the difference in the estimated treatment effect between cohorts 1 and 2 (i.e., $0.225 - 0.121 = 0.105$), and

[14]Results are similar when estimated using OLS.

Table 4: Dose-response Effects

	Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading compre-hension	Literacy score
	(1)	(2)	(3)	(4)	(5)
Days in tutorial x End of Grade 3	0.010***	0.002	0.005***	0.000	0.008***
	[0.002]	[0.001]	[0.001]	[0.001]	[0.001]
Days in tutorial x Beginning of Grade 4	0.007***	0.004***	0.006***	0.000	0.006***
	[0.001]	[0.001]	[0.002]	[0.001]	[0.002]
Days in tutorial x End of Grade 4	0.007***	0.001	0.001	0.003***	0.005***
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Observations	4949	6362	6362	6362	6362
p-value of equal coeffs.	0.244	0.017	0.005	0.016	0.239

Note: Each column shows the estimates of the coefficients β_h of equation (3), that is, the estimated dose response at different time horizons for each outcome of interest. Dosage is measured by the number of days in which students attended the tutorial. The actual attendance was instrumented with the randomized treatment indicator variable. The row labeled 'p-value of equal coeffs.' shows the p-value of a test $H_0 : \beta_1 = \beta_2 = \beta_3$. The average and standard deviation of the outcomes of students in the control group at baseline can be seen in Table 2. All models include cohort, year, and strata 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%.

about a third of the increase between cohorts 1 and 3 (i.e., $0.609 - 0.121 = 0.488$).

5.2 Targeting

A second change introduced after cohort 1 was to use fluency of reading rather than a composite literacy score as our eligibility variable. We also changed the traditional EGRA 60-words reading subtask for a longer 132-word text.[15] In order to assess whether the students deemed eligible changed over time, we compare the performance of eligible students at baseline using three subtasks that were identical in the three data-collection exercises: reading of non-words, addition, and subtraction. Table 5 shows the average differences in performance (not standardized) on these outcomes. We find that students in cohorts 2 and 3 had lower levels of skills than those eligible in cohort 1, and that students in the third cohort were better than those in the second cohort.

If the impact of the intervention is heterogeneous on students' skill levels, this may help to explain the different impacts observed between the three cohorts. Figure 6 investigates this by estimating quantile treatment effects following Firpo [2007].[16] We find that the treatment effect on knowledge of letter sounds is larger at the top quantiles. However, the relationship is flat for the composite literacy score and for the other subtasks. Therefore, it seems unlikely that the improvement in effectiveness was driven by the weaker set of students targeted in

[15] In addition, for cost reasons, we eliminated from the experimental sample schools with only one eligible student. Because these schools contribute with very few observations to the estimation, results do not change if we drop them.

[16] For an application of the estimation of quantile treatment effects, see, for instance, Bitler et al. [2017].

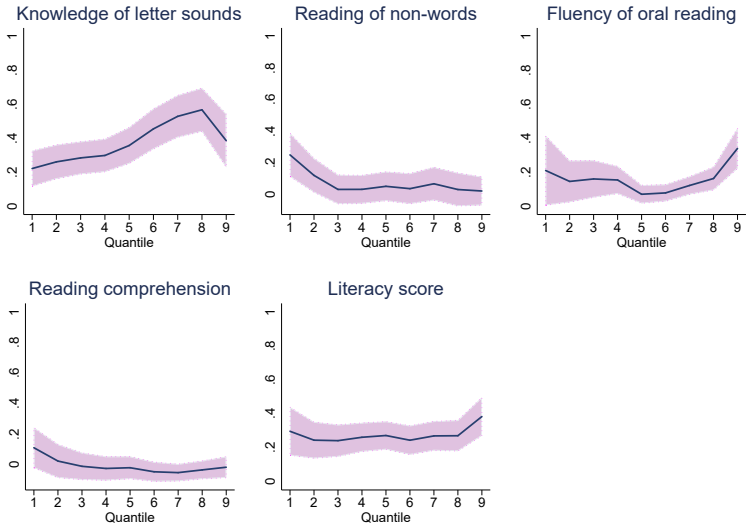
Table 5: Targeting

	Non-words	Addition	Subtraction	Index
	(1)	(2)	(3)	(4)
Cohort 2	-4.024***	-2.151***	-1.425***	-2.533***
	[0.718]	[0.296]	[0.267]	[0.348]
Cohort 3	-1.848***	-0.375	-0.057	-0.760**
	[0.704]	[0.291]	[0.282]	[0.355]
Observations	2610	2610	2610	2610
p-value of equal coeffs.	0.003	0.000	0.000	0.000

Note: Each column shows an OLS estimate of a model in which the dependent variable is an outcome measured at baseline (measured by the number of correct answers in that subtask) and the independent variables are dichotomous variables indicating that the students belong to cohort 2 or cohort 3. The index showed in column 4 is a simple average of the scores in the three subtasks shown in columns 1-3. All models include cohort, year, and strata 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%.

the last two cohorts of the experiment.[17]

Figure 6: Quantile Treatment Effects



Note: Each panel shows the quantile treatment effects on each outcome of interest estimated following Firpo [2007].

[17]Results available from the authors upon request show that, consistent with the quantile estimates, interacting the treatment variable with the baseline index of skills we use in Table 1 produces interactions effects that are small in magnitude, and we cannot reject that they are equal to zero.

5.3 Tutorial Composition

A third change addressed the composition of the students who attended tutorials. In the first cohort of the experiment we allowed the NGO to assign students to tutorials based on logistical considerations. Furthermore, to allow more students to benefit from the tutorials, the cut-off for the first cohort varied by school to accommodate as many students as possible – that is, up to the maximum of six per tutorial. In cohorts 2 and 3, we eliminated any discretion by randomizing each eligible student to a tutorial (in schools with more than one tutorial), and by using the same eligibility rule in all schools, regardless of the effect this had on tutorial sizes. As can be seen in Table 1, the tutorial size was on average 5.9 in cohort 1 and only 4.8 in cohorts 2 and 3.[18]

We investigate whether these changes in tutorial composition can partly explain the differential impact across cohorts by estimating intention-to-treat effects for students attending tutorials of different characteristics. Table 6 shows the treatment effects for two groups and the p-value of the test of equality of those effects. Contrary to what we were expecting, the first panel shows that larger tutorials were more effective at improving students' outcomes. Students in tutorials populated with six students did better in all subtasks than students in smaller tutorials.[19] The composition itself did not seem to make a difference in performance. We characterize the distribution of peers' ability by looking at an index based on a set of subtasks that are comparable across cohorts (i.e., reading of non-words, addition and subtraction). For each student we compute the mean of that index at baseline and checked whether it falls above or below the median. Students sitting with higher-ability peers performed similarly to those sitting with lower-ability peers. We also study the difference in performance of students sitting in more homogeneous versus heterogenous tutorials, again based on an index of comparable subtasks measured at baseline. More homogeneous groups tended to perform better, but the differences are not statistically significant at normal levels. Taken together, these results suggest that neither the size nor the composition of the tutorial groups can explain the increasing effectiveness of the intervention over time.

5.4 Tutors' Experience

About 40 percent of tutors in cohorts 2 and 3 had taught in a previous cohort. In cohort 2 the share of students taught by a tutor with previous experience was 0.45 while in cohort 3 this share was 0.33. The last rows of Table 6 present the learning gains of students taught by tutors with or without previous experience. As tutors in cohort 1 had no experience, we estimate these last two columns using cohorts 2 and 3. We find that students that received instructions from experienced tutors gained 0.15 of a standard deviation more in the overall literacy score. However, this difference is not statistically significant. These differential impacts in the overall score, mask some heterogeneity across subtasks. Students of less experienced tutors did better in knowledge of letter sounds, while students of more experienced tutors fared better in reading. This may reflect differences in allocation of time

[18] Of course, this also allowed more able students into the tutorials in cohort 1. However, as Figure 6 shows, this aspect of heterogeneity does not seem to offer an explanation for the gains we observe over time.

[19] Students are classified according to the *observed* number of students in the tutorial.

Table 6: Treatment Effect Heterogeneity

		Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading comprehension	Literacy score
		(1)	(2)	(3)	(4)	(5)
Size of tutorial group	1-5 students	0.336***	0.002	0.110*	-0.041	0.202***
		[0.064]	[0.053]	[0.058]	[0.051]	[0.060]
	6 students	0.378***	0.153**	0.280***	0.046	0.370***
		[0.082]	[0.059]	[0.060]	[0.033]	[0.072]
	p-value of equal coeffs.	0.662	0.016	0.029	0.141	0.038
Peers' initial ability	High	0.368***	0.065	0.102	-0.045	0.233***
		[0.067]	[0.062]	[0.064]	[0.044]	[0.067]
	Low	0.344***	0.086	0.286***	0.048	0.336***
		[0.071]	[0.063]	[0.067]	[0.042]	[0.072]
	p-value of equal coeffs.	0.765	0.803	0.051	0.108	0.253
Homogeneity of tutorial group	s.d. below median	0.379***	0.102*	0.233***	0.028	0.326***
		[0.066]	[0.055]	[0.060]	[0.039]	[0.065]
	s.d. above median	0.328***	0.044	0.145**	-0.031	0.233***
		[0.066]	[0.059]	[0.059]	[0.041]	[0.063]
	p-value of equal coeffs.	0.462	0.371	0.258	0.235	0.206
Tutor's previous experience	No	0.488***	0.032	0.239***	0.034	0.353***
		[0.102]	[0.063]	[0.072]	[0.042]	[0.085]
	Yes	0.256***	0.250***	0.556***	0.082	0.504***
		[0.096]	[0.077]	[0.120]	[0.081]	[0.103]
	p-value of equal coeffs.	0.056	0.016	0.018	0.570	0.178

Note: Each panel shows estimates of θ_1 of equation (2) estimated separately for two different groups of treated students where the comparison is against all the students in the control group. The row 'p-value of equal coeffs.' shows the p-value of a Chow-test of equality of coefficients in the different samples. See text for the description of each dimension of heterogeneity that is explored in the table. All models include cohort, year, and strata 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%.

to different activities between more and less experienced tutors.

5.5 Fine-Tuning of Material

So far we have explored quantifiable changes that could explain the increased effectiveness of the intervention over time. The analysis suggests that the increased dosage played an important role. Other factors such as the targeting of the intervention, the composition of the tutorial groups, and the increased experience of the tutors seem less important.

A last factor, more difficult to quantify, is that some of the difference might be also attributed to the fine-tuning of the material that occurred from cohorts 1, to 2, to 3. A first-order modification between the first and subsequent cohorts dealt with adjusting the difficulty of the texts used. Text difficulty is a key factor for comprehension. Texts that are too easy do not challenge students by providing enough difficult words. Texts that are too difficult do not provide enough opportunities to practice fluency, and may prevent the activation of complex

processes of comprehension.[20] In addition to these changes we included warm-up phonological awareness exercises, reorganized the readings in some of the sessions, and replaced some exercises related to vocabulary development in favor of others that further promoted reading fluency.

Unfortunately, it is not possible to quantify how much this could have contributed to the gains. The only variations between cohorts 2 and 3 are the tutors' experience and the fine-tuning of the material. Thus, we speculate that the adjustment of materials is one of the main drivers of the differences in the effectiveness of the intervention between the second and third cohorts.

6 Cost-Effectiveness

A natural comparison with our evaluation of a tutoring remediation program is the "Balsakhi" program analyzed by Banerjee et al. [2007] implemented in India in 2001. The authors find an average learning gain of 0.28 standard deviation (σ) at a cost of USD 2.25 per student. The tutoring intervention analyzed in this paper is similar in terms of effectiveness, with students gaining 0.286σ with a cost of implementing the intervention of USD 89 per student in 2016. The largest items driving the cost were wages and transportation of tutors.[21] To compare both costs, we can translate them into a common unit: the "Balsakhi" tutoring program translates into a cost of 0.5 percent of forgone consumption per capita, while the intervention evaluated in this paper achieves a similar learning gain but costs 1.5 percent of forgone consumption per capita.[22] This difference in cost is likely explained by economies of scale. While our tutorials had up to six students, those evaluated in Banerjee et al. [2007] had 15 to 20 children. The costs of our intervention are likely to be smaller in larger school districts where transportation costs are lower and tutors could teach more children per day (by offering more sessions).

A second way to generate a policy-relevant indicator of cost-effectiveness is to compare learning gains and costs during the relevant school year with those of the intervention itself. Third-grade students in the control group increased learning by 0.18σ per 100 dollars spent, whereas our intervention achieved a learning a gain of 0.32σ per 100 dollars. However, as noted by Muralidharan et al. [2019], while spending in education can increase unboundedly over time, students are in school only five hours a day. For this reason, evaluating the effectiveness of the program in terms of its time costs is also important. Again, students in third grade gained about 0.12σ per 100 hours of class, while students in our tutorials gained 0.89σ per 100 hours.[23]

[20] Students are trying to decode words whose meaning they do not know. Texts that are too easy do not provide enough opportunity to practice more difficult words. Beach and O'Connor [2014] argue for a potential threshold effect: it is necessary to select texts in which students can read at least 85 percent of words accurately to foster meaningful fluency growth.

[21] See Appendix Table A.3

[22] According to the World Development Indicators, GPD per capita in current dollars was 451 for India in 2001 and 5,871 for Colombia in 2016. Thus, $2.25/451 \times 100 = 0.5$ and $89/5871 \times 100 = 1.5$

[23] Recall from Section 4 that students in our program gained 0.286σ, and that during third grade, students

7 Conclusion

In countries where many students are reading below grade level in elementary school, it is important to find effective remediation methods so that students acquire basic skills that they need to progress in school and in life. We present the results of a remedial tutorial program conducted for small groups of third-grade students who are struggling to read. Outside school instructors followed a structured curriculum to implement a 16-week remediation program during school hours three times a week for 40 minutes. The experiment took place in the mid-size city of Manizales, Colombia, and involved 90 schools and more than 2,000 children in each of three different cohorts. Immediately after the experiment, reading fluency improved among treated children by no less than 20 percent of a standard deviation. We followed these children into the next academic year where these gains persist. We find that the gains of the program increased for each subsequent cohort that received the program.

Duflo [2017] argues that, in designing successful policies, economists should view their work like that of plumbers. They "will use a number of things...to tune every feature of the policy as well as possible, keeping an eye on all the relevant details as best he can. But with respect to some details, there will remain genuine uncertainty about the best way to proceed" (p. 4). Some of this uncertainty can be resolved by learning through experimentation. Our paper offers a good example of how economists can take this approach by using sequential experiments to adapt, refine, and test design features of a policy to beneficial effect. In our first experimental cohort we found limited gains to the intervention. In conversation with our partners we decided to change several issues to address potential factors that explained our limited success initially. These steps included targeting, tutorial composition, dosage and the design of the material. On the one hand, by continuing experimentation with subsequent cohorts we were able to show that increasing dosage (i.e., by offering more sessions and make-up sessions) and material design are important in explaining the gains we observe over time. On the other hand, we showed that the results are homogeneous across the ability distribution, and that changes in tutorial size and composition are not important factors in explaining the success of subsequent interventions.

We take our intervention to be a cost-effective remediation program. However, the results of the paper should not to be interpreted as arguing against *earlier* interventions. Indeed, taking similar steps earlier could be even more cost-effective. A prime intervention technique could be changing the way reading is taught in earlier grades of school so that fewer children reach the third grade still struggling to read.

in the control group gained 0.4σ. Students in third grade spend 1,000 hours in class at an annual cost of USD 665 per student (OECD [2019]). We assume, conservatively, that students spend one-third of the time in class acquiring literacy skills. Students in our program spent a total of 32 hours in the tutorials at a cost of USD 89 per student. This yield a per USD 100 effect of $(100 \times 0.40)/(1000/3) = 0.12$ for students in the control group and $(100 \times 0.286)/32 = 0.89$. Similarly, this yields a per 100 hours effect of $(100 \times 0.4)/(665/3) = 0.18$ for students in the control group and $(100 \times 0.286/89 = 0.32)$.

References

Abeberese, A. B., Kumler, T. J. and Linden, L. L. [2014], 'Improving reading skills by encouraging children to read in school: A randomized evaluation of the sa aklat sisikat reading program in the Philippines', *Journal of Human Resources* **49**(3), 611–633.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M. and Walton, M. [2017], 'From proof of concept to scalable policies: Challenges and solutions, with an application', *Journal of Economic Perspectives* **31**(4), 73–102.

Banerjee, A. V., Cole, S., Duflo, E. and Linden, L. [2007], 'Remedying education: Evidence from two randomized experiments in india*', *The Quarterly Journal of Economics* **122**(3), 1235–1264.

Banerji, R., Berry, J. and Shotland, M. [2017], 'The impact of maternal literacy and participation programs: Evidence from a randomized evaluation in india', *American Economic Journal: Applied Economics* **9**(4), 303–37.

Barrera-Osorio, F., Gonzalez, K., Lagos, F. and Deming, D. [2019], 'Effects, timing and heterogeneity of the provision of information in education: An experimental evaluation in colombia', *unpublished* .

Beach, K. and O'Connor, R. [2014], 'Developing and strengthening reading fluency and comprehension of poor readers in elementary school: A focused review of research.', *Perspectives on Language and Literacy* **40**(3).

Benavot, A. [1996], 'Education and political democratization: Cross-national and longitudinal findings', *Comparative Education Review* **40**, 377–403.

Bitler, M. P., Gelbach, J. B. and Hoynes, H. W. [2017], 'Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? evidence from a social experiment', *The Review of Economics and Statistics* **99**(4), 683–697.

Borkum, E., He, F. and Linden, L. L. [2012], 'The effects of school libraries on language skills: Evidence from a randomized controlled trial in india', *NBER Working Paper* (18183).

Busso, M., Cristia, J., Hincapie, D., Julian, M. and Ripani, L. [2017], *Learning Better: Public Policy for Skills Development*, Development in the Americas, IDB Publications, chapter 3, pp. 45–68.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. and Yagan, D. [2011], 'How does your kindergarten classroom affect your earnings? evidence from Project STAR', *The Quarterly Journal of Economics* **126**(4), 1593–1660.

Cunha, F. and Heckman, J. [2007], 'The technology of skill formation', *American Economic Review* **97**(2), 31–47.

Currie, J. and Thomas, D. [1995], 'Does Head Start make a difference?', *The American Economic Review* **85**(3), 341–364.

Deming, D. [2009], 'Early childhood intervention and life-cycle skill development: Evidence from Head Start', *American Economic Journal: Applied Economics* **1**(3), 111–34.

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

Duflo, E. [2017], 'Richard T. Ely Lecture: The economist as plumber', *American Economic Review* **107**(5), 1–26.

Ehri, L. [2005], 'Learning to read words: Theory, findings, and issues', *Scientific Studies of Reading* **9**, 167–188.

Elbaum, B., Vaughn, S., Tejero, H. and Moody, S. W. [2000], 'How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? a meta-analysis of the intervention research.', *Journal of Educational Psychology* **92**(4), 605–619.

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

Foorman, B. and 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.

Fuchs, L. S., Fuchs, D., Hosp, M. K. and Jenkins, J. R. [2001], 'Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis', *Scientific Studies of Reading* **5**(3), 239–256.

Glewwe, P., Kremer, M. and Moulin, S. [2009], 'Many children left behind? textbooks and test scores in Kenya', *American Economic Journal: Applied Economics* **1**(1), 112–35.

Good, R., Simmons, D. and Kameenui, E. [2001], 'The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes', *Scientific Studies of Reading* **5**, 257–288.

Good, R., Simmons, D. and Smith, S. [1998], 'Effective academic interventions in the united states: Evaluating and enhancing the acquisition of early reading skills.', *School Psychology Review* **27**(1), 45–56.

Goux, D., Gurgand, M. and Maurin, E. [2017], 'Reading enjoyment and reading skills: Lessons from an experiment with first grade children', *Labour Economics* **45**, 17 – 25.

Hanushek, E. A., Schwerdt, G., Wiederhold, S. and Woessmann, L. [2015], 'Returns to skills around the world: Evidence from piaac', *European Economic Review* **73**, 103 – 130.

Hart, B. and Risley, T. R. [1995], *Meaningful differences in the everyday experience of young American children.*, Paul H Brookes Publishing.

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

Inns, A., Lake, C., Pellegrini, M. and Slavin, R. [2019], 'A quantitative synthesis of research on programs for struggling readers in elementary schools.', *Best Evidence Encyclopedia, Center for Research and Reform in Education.* .

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. and Thornton, R. [2019], 'Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures', *unpublished* .

Machin, S. and McNally, S. [2008], 'The literacy hour', *Journal of Public Economics* **92**(5), 1441 – 1462.

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

Muralidharan, K., Singh, A. and 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 [2016], *Skills Matter: Further Results from the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris.

OECD [2019], *Education at a Glance 2019*.

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

Sentell, T. L. and Halpin, H. A. [2006], 'Importance of adult literacy in understanding health disparities', *Journal of general internal medicine* **21**, 862–866.

Shaywitz, S. E. and Shaywitz, B. A. [1996], *Unlocking learning disabilities: The neurological basis*.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A. and Davis, S. [2009], 'Effective reading programs for the elementary grades: A best-evidence synthesis', *Review of Educational Research* **79**(4).

Stanovich, K. [1986], 'Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy', *Reading Research Quarterly* **22**, 360–407.

UNESCO [2005], *Education for All: Literacy for Life*, United Nations Educational Scientific and Cultural Organization.

Whitehurst, G. J. and Lonigan, C. J. [1998], 'Child development and emergent literacy', *Child development* **69**(3), 848–872.

Zhang, X., Koponen, T., Räsänen, P., Aunola, K., Lerkkanen, M.-K. and Nurmi, J.-E. [2014], 'Linguistic and spatial skills predict early arithmetic development via counting sequence knowledge', *Child Development* **85**(3), 1091–1107.

A Appendix Tables and Figures

Table A.1: Treatment Effects on all outcomes and all time horizons

Experimental cohort	Moment in time	Outcome	Treatment effect	Standard error
Cohort 1	End of Grade 3	Knowledge of letter sounds	0.301	[0.081]***
		Reading of non-words	0.024	[0.064]
		Fluency of oral reading	-0.025	[0.041]
		Reading comprehension	-0.064	[0.041]
		Literacy score	0.120	[0.060]**
	Beg/Med. Grade 4	Knowledge of letter sounds	-	-
		Reading of non-words	-0.000	[0.079]
		Fluency of oral reading	0.007	[0.119]
		Reading comprehension	0.196	[0.063]***
		Literacy score	0.016	[0.110]
	End of Grade 4	Knowledge of letter sounds	-	-
		Reading of non-words	0.098	[0.049]**
		Fluency of oral reading	-0.012	[0.067]
		Reading comprehension	-0.032	[0.043]
		Literacy score	0.017	[0.057]
Cohort 2	End of Grade 3	Knowledge of letter sounds	0.137	[0.104]
		Reading of non-words	0.053	[0.080]
		Fluency of oral reading	0.266	[0.082]***
		Reading comprehension	0.048	[0.055]
		Literacy score	0.225	[0.082]***
	Beg/Med. Grade 4	Knowledge of letter sounds	0.171	[0.076]**
		Reading of non-words	0.049	[0.080]
		Fluency of oral reading	0.326	[0.101]***
		Reading comprehension	0.074	[0.068]
		Literacy score	0.338	[0.095]***
	End of Grade 4	Knowledge of letter sounds	0.234	[0.071]***
		Reading of non-words	0.027	[0.084]
		Fluency of oral reading	0.061	[0.065]
		Reading comprehension	0.062	[0.029]**
		Literacy score	0.156	[0.081]*
Cohort 3	End of Grade 3	Knowledge of letter sounds	0.678	[0.115]***
		Reading of non-words	0.183	[0.076]**
		Fluency of oral reading	0.462	[0.108]***
		Reading comprehension	0.056	[0.059]
		Literacy score	0.609	[0.109]***
	Beg/Med. Grade 4	Knowledge of letter sounds	0.487	[0.100]***
		Reading of non-words	0.191	[0.081]**
		Fluency of oral reading	0.205	[0.088]**
		Reading comprehension	0.161	[0.079]**
		Literacy score	0.405	[0.107]***
	End of Grade 4	Knowledge of letter sounds	0.389	[0.086]***
		Reading of non-words	0.131	[0.081]
		Fluency of oral reading	0.185	[0.100]*
		Reading comprehension	0.137	[0.102]
		Literacy score	0.343	[0.104]***

Note: Each panel shows separate estimates of the treatment effects for each cohort. We estimate the models separately for each time-horizon and for each outcome of interest. All models include strata fixed effects. We did not collected data on the outcomes without an estimated treatment effect (marked as -). 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.2: Treatment Effects Robustness

		Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading comprehension	Literacy score
		(1)	(2)	(3)	(4)	(5)
Control: Baseline outcome (Y0)	Treatment x End of Grade 3	0.348***	0.072	0.202***	-0.002	0.286***
		[0.055]	[0.045]	[0.050]	[0.032]	[0.052]
	Treatment x Beginning of Grade 4	0.333***	0.133**	0.198**	0.104**	0.280***
		[0.063]	[0.057]	[0.076]	[0.050]	[0.074]
	Treatment x End of Grade 4	0.309***	0.078*	0.093*	0.033	0.159***
		[0.048]	[0.041]	[0.052]	[0.039]	[0.048]
Controls: Individual and School	Treatment x End of Grade 3	0.357***	0.077*	0.199***	0.007	0.290***
		[0.059]	[0.046]	[0.048]	[0.034]	[0.054]
	Treatment x Beginning of Grade 4	0.322***	0.105**	0.190**	0.118**	0.273***
		[0.069]	[0.050]	[0.076]	[0.050]	[0.075]
	Treatment x End of Grade 4	0.300***	0.090**	0.092*	0.044	0.169***
		[0.056]	[0.039]	[0.050]	[0.041]	[0.049]
School fixed effects	Treatment x End of Grade 3	0.456***	0.117	0.186**	0.040	0.328***
		[0.097]	[0.085]	[0.084]	[0.061]	[0.097]
	Treatment x Beginning of Grade 4	0.407***	0.152*	0.166*	0.143**	0.301***
		[0.090]	[0.087]	[0.100]	[0.067]	[0.104]
	Treatment x End of Grade 4	0.384***	0.132	0.084	0.075	0.210**
		[0.092]	[0.081]	[0.094]	[0.057]	[0.102]

Note: All models include cohort, year, and strata fixed effects. Models with controls include baseline outcome (panel 1) baseline outcome plus age, age square, gender, and disability status (panel 2) and all covariates with school fixed effects but no strata fixed effects (panel 3) which we can do because schools changed treatment status across cohorts. Each column-panel show the coefficients θ_h of equation (2). That is, the estimated treatment effects at different time horizons for each outcome of interest. All models include cohort, year, and strata 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%.

Table A.3: Costs

		Cost in dollars
Cost per tutor		2158
Training	8 hours. Includes cost of 1 trainer per 15 tutors	40
Wages and transport	Includes wages and transport	2067
Supervision	Supervision of 2 classes per tutor. Includes salaries and transport cost	51
Cost per student	Students per tutor = 26	89
Materials		6
Tutor		83

Note. Parameters: Tutorials per tutor 5; Number of tutors 15; Exchange rate 3,000 Colombian pesos per dollar of 2016; Total number of students in tutorials 385; Total number of tutorials 76; Hourly wage 4.167.

B Data Appendix

This paper relies on two sources of information. First, before the intervention, we administered language and mathematics tests to third grade students at the beginning of each year (2015, 2016 and 2017) in all public schools in the Municipality of Manizales. After that, to measure the impact of the intervention, we collected information in the same schools at the end of Grade 3, the beginning/middle and the end of Grade 4 using the same instrument as well. The former data consist 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.2.

At the time of baseline of each school year, 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.1 Sample Sizes

In Table B.1 we show the number of schools, classrooms and students that participated in the experiment. We started off with a sample of 94 public schools in 2015, which we dropped, mostly for cost reasons, to 84 and then 80 in the two subsequent years. The number of third grade classrooms within these schools ranged between 124 and 155. A total of 1,143, 753

Table B.1: Sample Sizes

	Schools		Classrooms		Students	
Cohort	Control	Treated	Control	Treated	Control	Treated
1	46	48	74	75	591	552
2	42	42	57	61	377	376
3	40	40	60	62	353	361

and 714 students were eligible in cohorts 1, 2 and 3. It is important to note that when we measure the outcomes during Grade 4 we also follow up any retained students in third grade who initially were in our sample either in treatment or control groups but have not been promoted to Grade 4 during the next academic year.

B.2 Attrition

We were not able to follow up all the students from our baseline sample in the subsequent times that we visited the schools. We estimate the probability of attrition as a function the baseline standardized scores for knowledge of letter sounds, reading of non-words, fluency of oral reading and reading Comprehension. The following table shows that, in general, the baseline scores of students that attrited from our sample are similar to the scores of those who attrited, suggesting that there is no clear pattern between the characteristics at baseline and attrition across study arms. Most estimates are small. We only reject the null of zero difference in two cases, in different directions and at the 10 percent level.

Table B.2: Attrition

	End of Grade 3	Beginning of Grade 4	End of Grade 4
	(1)	(2)	(3)
Knowledge of letter sounds	-0.024	0.010	-0.048*
	[0.006]	[0.010]	[0.009]
Reading of non-words	-0.018	-0.010	-0.012
	[0.009]	[0.011]	[0.011]
Fluency of oral reading	-0.043	0.053*	-0.029
	[0.009]	[0.011]	[0.010]
Reading comprehension	0.017	-0.024	-0.008
	[0.009]	[0.010]	[0.010]
Observations	2,610	2,610	2,610

Note: Standard errors, shown in squared brackets, are clustered at the school level. * significant at 10%; ** significant at 5%; *** significant at 1%.

Instruments and Test Scores

As described in Section 3.1 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 two subtasks of math: knowledge of letter sounds, reading of non-words,

fluency of oral reading, reading comprehension, addition and subtraction. In the first two components, children are asked to recognize the letters and 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 two subtasks involve students solving math operation of one- and two- digit numbers to measure their math knowledge. We 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-cohort 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.

Test Subtasks Over Time

Table B.3 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 subtasks. Even though some minor scaling differences may exist, the outcomes are comparable across the different grades and cohorts because the items have been modified in order to be as similar as possible in terms of length and difficulty.

Table B.3: Scales by Grade and Cohort

	Cohort 1				Cohort 2				Cohort 3			
Grade:	3	3	4	4	3	3	4	4	3	3	4	4
Time of School Year:	Beg.	End	Beg.	End	Beg.	End	Beg.	End	Beg.	End	Beg.	End
Scales:												
Knowledge of Letter Sounds	1	1	-	-	1	1	2	2	2	2	2	3
Reading Fluency	1	2	3	4	3	5	5	5	5	5	4	4
Reading Comprehension	1	2	3	4	5	6	7	7	7	7	8	8
Reading of Non-words, Addition, Substraction	Always the same scales in all cohorts and years											

Table B.3 presents all the items for all the subtasks and tests. Column 1 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 this subtask, as well as reading comprehension, is not constructed with a constant number of items across instruments. For example, the first scale contains 59 words while the second one has 64 words in the paragraph. However, despite these differences, the scores are comparable since we created a composite score by standardizing them for each grade, year and cohort to place students on the same scale. Lastly, the scales were exactly the same for all cohorts and years for the reading of non-words, addition and subtraction subtasks.

Item Response Theory

Item response theory (IRT) is used in the design, analysis, scoring, and comparison of tests and similar instruments whose purpose is to measure unobservable characteristics of the respondents. IRT models specify a relationship between a single underlying latent achievement variable (ability) and the probability of answering a particular test question (item) correctly. We use the Rasch model to assess items and to score subjects on their abilities or other latent traits. In the Rasch model, the probability of a correct response is given by

$$Pr\left(Y_{ij} = 1|\theta_j\right) = \frac{e^{\alpha(\theta_j - b_i)}}{1 + e^{\alpha(\theta_j - b_i)}}$$

where α represents the discrimination common to all items, b_i represents the difficulty of item i, and θ is the latent trait of person j. The probability of a correct response is determined by the item's difficulty and the subject's ability.

Appendix Tables B.6 - B.9 presents a detailed item analysis of four subtasks administered at the endline of Grade 3 to ensure that the items performed well in terms of discrimination and difficulty. We also assess how reliable the scale is for each subtask by computing Cronbach's alpha coefficient. All alpha coefficients of all four different scales are higher than 0.8, suggesting that the items have relatively high internal consistency across study arms.

Table B.4: Intra-Item Correlation Matrix

Subtask	Time of School Year	Baseline	Endline G3	Beg/Mid. G4	Endline G4
Knowledge of letter sounds	Baseline	1.000			
	Endline G3	0.425	1.000		
	Beg/Mid. G4	0.365	0.510	1.000	
	Endline G4	0.425	0.533	0.518	1.000
Reading of non-words	Baseline	1.000			
	Endline G3	0.581	1.000		
	Beg/Mid. G4	0.479	0.560	1.000	
	Endline G4	0.512	0.580	0.585	1.000
Reading Fluency	Baseline	1.000			
	Endline G3	0.693	1.000		
	Beg/Mid. G4	0.305	0.397	1.000	
	Endline G4	0.492	0.585	0.515	1.000
Reading Comprehension	Baseline	1.000			
	Endline G3	0.231	1.000		
	Beg/Mid. G4	0.135	0.163	1.000	
	Endline G4	0.076	0.162	0.317	1.000
Addition	Baseline	1.000			
	Endline G3	0.598	1.000		
	Beg/Mid. G4	0.465	0.562	1.000	
	Endline G4	0.525	0.591	0.570	1.000
Substraction	Baseline	1.000			
	Endline G3	0.467	1.000		
	Beg/Mid. G4	0.319	0.414	1.000	
	Endline G4	0.433	0.493	0.432	1.000

Note: All values are significant at 1%.

Table B.5: Test items

Knowledge of Letter Sounds		Reading Fluency		Reading Comprehension		Reading of Non-words		Addition		Substraction	
ID	Scale	ID	Scale	ID	Scale	ID	Scale	ID	Scale	ID	Scale
1	M d r O E C i u p S 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	1	María y Juan fueron a jugar al parque. En el parque estaba Jaime. Jaime había perdido a su peluche. Estaba muy triste. María y Juan lo ayudaron a buscar. El peluche estaba dormido bajo un árbol. Despierta, peluche, vamos a jugar dijo Jaime. Los niños jugaron hasta que llegó la noche. Todos estaban contentos ese día. El peluche también.	1	¿A dónde fueron a jugar María y Juan? ¿Por qué estaba triste Jaime? ¿Qué hicieron Juan y María al ver triste a Jaime? ¿Dónde estaba el peluche? ¿Hasta cuándo jugaron los niños en el parque?	1	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	1	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)	1	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	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	2	Había un perrito gordo y peludo llamado Toto. La familia con quien vivía lo quería mucho. Toto era un perro obediente, cuidaba muy bien la casa, pero no se comía toda su comida. Un día salió de paseo con su dueño Lucas y se perdió. Lucas se puso triste, pero felizmente Todo apareció en el parque. Lucas lo cargó y lo llevó a casa.	2	¿Cómo se llamaba el perro? ¿Qué hacía muy bien el perro? ¿El perro es flaco o gordo? ¿Dónde apareció el perro? ¿Quién es Lucas?						
3	M d r O E C i u p S 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	3	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.	3	¿Qué se pusieron a jugar la gallina y el cienpiés? ¿Por qué el cienpiés pudo atrapar muchas pelotas? ¿Cuántos goles anotó el cienpiés? ¿Por qué la gallina estaba furiosa? ¿Qué hizo la gallina con el cienpiés?						
		4	Pedro y Mateo son amigos. A ambos les gusta jugar fútbol. Les gusta ir a la escuela, entre otras cosas, a jugar pelota a la hora del recreo. Forman equipos, algunas veces por grado, otras por afinidad. El recreo se siente cortico porque jugar pelota es entretenido. Los niños corren, saltan, patean y gritan. Lo mejor es cuando su equipo mete gol. La mayoría de veces Miguel lleva el balón para jugar el partido. El viernes pasado Miguel no llegó a la escuela porque tenía varicela. Pedro pensó que sería un recreo aburrido, pues no habría partido. Mateo tuvo una idea brillante. Buscaron hojas de papel usadas. Preguntaron al maestro si se las regalaba. Él amablemente accedió. Arrugaron una hoja e hicieron una bola. Luego pusieron más hojas hasta hacer una gran pelota. El maestro los vio y los ayudó. Puso cinta adhesiva a su pelota de papel. La idea del maestro era que la pelota durara por más tiempo, hasta que regresara Miguel.	4	¿Qué les gusta jugar a Pedro y a Mateo? ¿Quién lleva el balón para jugar el partido? ¿Qué día de la semana faltó Miguel a la escuela? ¿Por qué no fue Miguel a la escuela? ¿Por qué pensó Pedro que el recreo sería aburrido?						
		5	El abuelo tomaba café. Era una tarde lluviosa. Recordaba cuando era niño. El abuelo contó, como era la siembra 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: -¿cómo 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. Después se tuesta y 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.	5	¿Qué se pusieron a jugar la gallina y el cienpiés? ¿Quién pateó más lejos? ¿Por qué el cienpiés pudo atrapar muchas pelotas? ¿Cuántos goles anotó el cienpiés? ¿Por qué la gallina estaba furiosa?						
				6	¿Cómo estaba la tarde? ¿Cómo era el pueblo? ¿Cómo se llamaba el pueblo? ¿A qué ayudaba el abuelo cuando era niño? ¿De qué color es el fruto que da el Cafeto?						
				7	¿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?						
				8	¿Qué les gusta jugar a Pedro y a Mateo? ¿Cómo se forman los grupos para jugar fútbol? ¿Cómo se siente el recreo cuando juegan pelota? ¿En qué momento les gusta jugar fútbol a Pedro y a Mateo? ¿Quién lleva el balón para jugar el partido? ¿Por qué no fue Miguel a la escuela? ¿Por qué pensó Pedro que el recreo sería aburrido? ¿Cuál fue la idea brillante de Mateo?						

Table B.6: IRT of Knowledge of Letter Sounds

		Cohort 1			Cohort 2			Cohort 3	
Knowledge of letter sounds	Item	Difficulty Parameter	Std. Error	Item	Difficulty Parameter	Std. Error	Item	Difficulty Parameter	Std. Error
Item1	M	-1.067	0.032	M	-0.980	0.030	M	-1.371	0.037
Item2	d	-0.552	0.027	d	-0.588	0.025	d	-0.734	0.027
Item3	r	-0.678	0.028	r	-0.728	0.027	r	-1.041	0.031
Item4	O	-1.077	0.032	O	-0.939	0.030	O	-1.142	0.033
Item5	E	-1.039	0.032	E	-0.966	0.030	E	-1.163	0.033
Item6	C	-0.729	0.028	C	-0.764	0.027	F	-0.907	0.029
Item7	i	-1.139	0.033	i	-0.981	0.030	i	-1.175	0.033
Item8	u	-1.050	0.032	u	-0.970	0.030	u	-1.152	0.033
Item9	p	-0.137	0.025	p	-0.242	0.023	p	-0.429	0.025
Item10	S	-0.545	0.027	S	-0.659	0.026	S	-1.132	0.032
Item11	A	-1.077	0.032	A	-0.872	0.029	A	-1.081	0.032
Item12	n	-0.305	0.025	n	-0.362	0.023	n	-0.746	0.027
Item13	j	0.099	0.025	j	-0.035	0.022	j	-0.171	0.024
Item14	T	-0.211	0.025	T	-0.206	0.023	T	-0.527	0.025
Item15	b	0.047	0.025	b	-0.057	0.022	b	-0.202	0.024
Item16	e	-0.795	0.029	e	-0.521	0.025	e	-0.727	0.027
Item17	f	-0.071	0.025	f	-0.155	0.022	f	-0.538	0.025
Item18	r	-0.125	0.025	r	-0.167	0.023	r	-0.444	0.025
Item19	W	0.912	0.030	W	0.600	0.025	G	-0.167	0.024
Item20	L	0.082	0.025	L	-0.014	0.022	L	-0.278	0.024
Item21	m	-0.132	0.025	m	-0.081	0.022	m	-0.426	0.025
Item22	R	0.009	0.025	R	0.010	0.022	R	-0.286	0.024
Item23	D	0.124	0.025	D	0.105	0.022	D	-0.142	0.024
Item24	E	-0.321	0.025	E	-0.013	0.022	E	-0.274	0.024
Item25	y	0.280	0.025	y	0.295	0.023	y	0.237	0.024
Item26	O	-0.234	0.025	O	0.100	0.022	O	-0.161	0.024
Item27	a	-0.186	0.025	a	0.135	0.022	a	-0.129	0.024
Item28	g	0.369	0.026	g	0.415	0.024	g	0.131	0.024
Item29	s	0.273	0.025	s	0.367	0.023	s	0.006	0.024
Item30	Z	0.423	0.026	Z	0.486	0.024	Z	0.178	0.024
Item31	c	0.336	0.025	c	0.503	0.024	f	0.196	0.024
Item32	V	0.657	0.027	V	0.681	0.026	V	0.480	0.026
Item33	N	0.558	0.027	N	0.677	0.026	N	0.336	0.025
Item34	I	0.251	0.025	I	0.600	0.025	I	0.319	0.025
Item35	k	0.715	0.028	k	0.855	0.028	b	0.551	0.026
Item36	U	0.349	0.025	U	0.714	0.026	U	0.415	0.025
Item37	P	0.815	0.029	P	0.953	0.029	P	0.639	0.027
Item38	x	1.037	0.031	x	1.079	0.031	R	0.625	0.027
Item39	L	0.931	0.030	L	1.080	0.031	L	0.699	0.028
Item40	Q	1.101	0.032	Q	1.172	0.033	M	0.701	0.028
Item41	S	0.963	0.030	S	1.127	0.032	S	0.738	0.028
Item42	Ñ	1.193	0.033	Ñ	1.310	0.035	v	0.906	0.030
Item43	O	0.780	0.029	O	1.101	0.031	O	0.842	0.029
Item44	A	0.826	0.029	A	1.137	0.032	A	0.890	0.030
Item45	d	1.219	0.034	d	1.349	0.036	d	1.039	0.031
Item46	T	1.317	0.035	T	1.413	0.037	T	1.120	0.032
Item47	i	0.989	0.031	i	1.286	0.035	i	1.086	0.032
Item48	N	1.394	0.036	N	1.476	0.038	N	1.171	0.033
Item49	a	1.064	0.032	a	1.326	0.035	a	1.213	0.034
Item50	e	1.100	0.032	e	1.359	0.036	e	1.297	0.035
Discrimination Parameter		2.521	0.036		3.709	0.058		3.527	0.055
Cronbach's Alpha	Beginning of Grade 3	0.957			0.963			0.969	
	End of Grade 3	0.965			0.972			0.974	
	End of Grade 4	N/A			0.973			0.976	
	Beginning of Grade 4	N/A			0.966			0.983	

Note: IRT estimates using a one-parameter model.

Table B.7: IRT of Reading of Non-words

Reading of non-words	Item	Cohort 1		Cohort 2		Cohort 3	
		Difficulty Parameter	Std. Error	Difficulty Parameter	Std. Error	Difficulty Parameter	Std. Error
Item1	lete	-2.253	0.059	-2.043	0.053	-2.036	0.055
Item2	quibe	-2.293	0.060	-2.049	0.054	-2.097	0.058
Item3	bofa	-1.778	0.044	-1.648	0.040	-1.723	0.045
Item4	mise	-2.594	0.073	-2.347	0.070	-2.269	0.065
Item5	garo	-1.415	0.037	-1.623	0.039	-1.535	0.040
Item6	cafa	-1.884	0.047	-1.695	0.041	-1.671	0.043
Item7	Celu	-2.620	0.075	-2.217	0.062	-2.140	0.059
Item8	bede	-0.960	0.031	-1.238	0.031	-1.047	0.032
Item9	lura	-2.046	0.052	-1.928	0.049	-1.866	0.049
Item10	mesi	-2.263	0.059	-2.192	0.061	-2.000	0.054
Item11	lluno	-2.158	0.055	-1.950	0.050	-2.041	0.055
Item12	Rite	-2.060	0.052	-1.994	0.051	-1.991	0.053
Item13	duso	-2.021	0.051	-2.021	0.052	-1.881	0.050
Item14	jata	-1.956	0.049	-1.822	0.045	-1.897	0.050
Item15	fica	-2.175	0.056	-2.049	0.054	-2.020	0.055
Item16	luma	-1.821	0.046	-1.756	0.043	-1.825	0.048
Item17	Alti	-2.197	0.057	-1.999	0.052	-1.940	0.052
Item18	lufa	-2.258	0.059	-2.010	0.052	-1.944	0.052
Item19	frate	-1.696	0.043	-1.643	0.040	-1.487	0.039
Item20	dulte	-1.859	0.046	-1.741	0.042	-1.679	0.043
Item21	ledo	-1.946	0.049	-1.811	0.044	-1.760	0.046
Item22	Fosu	-1.492	0.038	-1.602	0.039	-1.422	0.038
Item23	gesa	-1.323	0.036	-1.466	0.035	-1.278	0.035
Item24	lemo	-2.082	0.053	-1.695	0.041	-1.690	0.044
Item25	golpa	-1.778	0.044	-1.566	0.038	-1.524	0.040
Item26	bosa	-1.591	0.040	-1.397	0.034	-1.409	0.037
Item27	rale	-1.503	0.039	-1.342	0.033	-1.248	0.034
Item28	flano	-1.505	0.039	-1.295	0.032	-1.280	0.035
Item29	trabu	-1.407	0.037	-1.198	0.031	-1.201	0.034
Item30	bulo	-1.450	0.038	-1.184	0.030	-1.183	0.033
Item31	pluva	-1.160	0.033	-1.008	0.028	-1.005	0.031
Item32	arcu	-1.314	0.035	-1.031	0.028	-1.069	0.032
Item33	cince	-0.965	0.031	-0.837	0.027	-0.845	0.029
Item34	llusia	-1.049	0.032	-0.843	0.027	-0.835	0.029
Item35	firta	-0.979	0.031	-0.752	0.026	-0.705	0.028
Item36	onti	-0.785	0.029	-0.613	0.025	-0.613	0.028
Item37	zaca	-0.842	0.030	-0.587	0.025	-0.619	0.028
Item38	queno	-0.760	0.029	-0.531	0.025	-0.538	0.027
Item39	bana	-0.654	0.028	-0.394	0.024	-0.425	0.027
Item40	juru	-0.327	0.027	-0.187	0.024	-0.125	0.026
Item41	foba	-0.304	0.027	-0.145	0.024	-0.156	0.026
Item42	lise	-0.353	0.027	-0.136	0.024	-0.153	0.026
Item43	vodo	-0.219	0.027	-0.028	0.024	-0.048	0.026
Item44	tuzi	-0.158	0.027	0.046	0.024	0.016	0.026
Item45	listu	-0.034	0.027	0.141	0.025	0.157	0.027
Item46	quira	0.074	0.027	0.255	0.025	0.269	0.027
Item47	cuto	0.214	0.027	0.358	0.026	0.360	0.027
Item48	ganco	0.347	0.028	0.435	0.026	0.472	0.028
Item49	rafo	0.397	0.028	0.504	0.026	0.538	0.028
Item50	duba	0.491	0.028	0.592	0.027	0.696	0.030
Discrimination Parameter		2.174	0.033	3.386	0.056	2.820	0.046
Cronbach's Alpha	Beginning of Grade 3	0.952		0.958		0.957	
	End of Grade 3	0.943		0.954		0.955	
	Beginning of Grade 4	0.950		0.940		0.953	
	End of Grade 4	0.946		0.951		0.939	

Note: IRT estimates using a one-parameter model.

Table B.8: IRT of Addition

Addition	Item	Cohort 1 Difficulty Parameter	Cohort 1 Std. Error	Cohort 2 Difficulty Parameter	Cohort 2 Std. Error	Cohort 3 Difficulty Parameter	Cohort 3 Std. Error
Item1	2+2=(4)	-2.342	0.066	-2.352	0.075	-2.312	0.070
Item2	3+2=(5)	-1.401	0.034	-1.422	0.035	-1.590	0.040
Item3	4+2=(6)	-1.528	0.037	-1.517	0.037	-1.717	0.043
Item4	1+5=(6)	-1.850	0.045	-1.901	0.048	-2.033	0.055
Item5	3+4=(7)	-1.429	0.035	-1.449	0.035	-1.546	0.038
Item6	7+1=(8)	-1.785	0.043	-1.924	0.049	-2.021	0.054
Item7	6+2=(8)	-1.586	0.038	-1.684	0.041	-1.651	0.041
Item8	5+4=(9)	-1.567	0.038	-1.569	0.038	-1.560	0.039
Item9	4+5=(9)	-1.434	0.035	-1.412	0.035	-1.425	0.036
Item10	7+2=(9)	-1.284	0.032	-1.242	0.032	-1.289	0.033
Item11	6+4=(10)	-1.022	0.028	-0.968	0.028	-1.051	0.030
Item12	5+5=(10)	-1.089	0.029	-0.963	0.028	-1.043	0.030
Item13	8+2=(10)	-0.907	0.027	-0.785	0.026	-0.895	0.028
Item14	5+6=(11)	-0.695	0.025	-0.550	0.025	-0.628	0.026
Item15	6+6=(12)	-0.555	0.025	-0.356	0.024	-0.422	0.025
Item16	3+9=(12)	-0.333	0.024	-0.103	0.024	-0.187	0.025
Item17	5+7=(12)	-0.028	0.024	0.186	0.025	0.123	0.025
Item18	8+6=(14)	0.275	0.025	0.601	0.028	0.473	0.027
Item19	10+3=(13)	0.415	0.026	0.724	0.029	0.650	0.029
Item20	2+11=(13)	0.670	0.028	0.932	0.031	0.821	0.030
Item21	13+3=(16)	0.841	0.030	1.147	0.034	0.965	0.032
Item22	6+10=(16)	1.007	0.031	1.330	0.036	1.145	0.034
Item23	10+10=(20)	1.150	0.033	1.444	0.038	1.238	0.035
Item24	15+5=(20)	1.272	0.034	1.598	0.040	1.339	0.037
Item25	11+9=(20)	1.526	0.038	1.758	0.043	1.621	0.041
Discrimination Parameter		3.059	0.049	3.508	0.061	3.566	0.063
Cronbach's Alpha	Beginning of Grade 3	0.914		0.902		0.902	
	End of Grade 3	0.912		0.909		0.913	
	Beginning of Grade 4	0.917		0.886		0.921	
	End of Grade 4	0.908		0.909		0.909	

Note: IRT estimates using a one-parameter model.

Table B.9: IRT of Substraction

Substraction	Item	Cohort 1 Difficulty Parameter	Cohort 1 Std. Error	Cohort 2 Difficulty Parameter	Cohort 2 Std. Error	Cohort 3 Difficulty Parameter	Cohort 3 Std. Error
Item1	4-2=(2)	-1.969	0.049	-1.715	0.044	-1.922	0.050
Item2	8-1=(7)	-1.945	0.048	-1.815	0.047	-2.083	0.057
Item3	5-2=(3)	-1.755	0.043	-1.661	0.042	-1.854	0.048
Item4	6-2=(4)	-1.542	0.038	-1.525	0.038	-1.667	0.042
Item5	8-2=(6)	-1.405	0.035	-1.366	0.034	-1.496	0.038
Item6	6-5=(1)	-1.504	0.037	-1.456	0.036	-1.702	0.043
Item7	9-2=(7)	-1.277	0.033	-1.222	0.031	-1.423	0.036
Item8	9-4=(5)	-1.073	0.030	-1.061	0.029	-1.211	0.033
Item9	8-3=(5)	-0.925	0.028	-0.916	0.027	-0.982	0.030
Item10	9-5=(4)	-0.750	0.027	-0.678	0.025	-0.725	0.027
Item11	7-4=(3)	-0.465	0.025	-0.386	0.024	-0.476	0.026
Item12	10-2=(8)	-0.300	0.024	-0.172	0.024	-0.224	0.025
Item13	10-3=(7)	-0.025	0.024	0.102	0.024	0.051	0.026
Item14	10-4=(6)	0.186	0.025	0.333	0.026	0.317	0.027
Item15	20-10=(10)	0.369	0.025	0.539	0.027	0.545	0.028
Item16	11-6=(5)	0.743	0.028	0.904	0.030	0.887	0.031
Item17	11-7=(4)	1.048	0.031	1.213	0.034	1.258	0.035
Item18	12-9=(3)	1.246	0.033	1.556	0.039	1.553	0.040
Item19	12-7=(5)	1.555	0.038	1.838	0.045	1.846	0.046
Item20	12-6=(6)	1.775	0.043	2.066	0.050	2.083	0.053
Item21	13-11=(2)	1.999	0.048	2.309	0.058	2.270	0.059
Item22	14-6=(8)	2.210	0.054	2.516	0.065	2.547	0.071
Item23	16-3=(13)	2.335	0.058	2.675	0.071	2.751	0.081
Item24	16-10=(6)	2.504	0.064	2.977	0.085	2.830	0.086
Item25	20-5=(15)	2.695	0.072	3.179	0.096	2.919	0.091
Item26	20-4=(16)	2.761	0.075	3.250	0.100	2.939	0.092
Item27	20-9=(11)	2.861	0.080	3.447	0.113	3.067	0.100
Discrimination Parameter		2.890	0.045	3.710	0.066	3.255	0.057
Cronbach's Alpha	Beginning of Grade 3	0.900		0.864		0.882	
	End of Grade 3	0.895		0.891		0.884	
	Beginning of Grade 4	0.913		0.881		0.915	
	End of Grade 4	0.902		0.902		0.897	

Note: IRT estimates using a one-parameter model.