Remedial Education: Evidence from a Sequence of Experiments in Colombia*

Horacio Alvarez-Marinelli Samuel Berlinski Matias Busso

Abstract

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 instructors 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.

*Horacio Alvarez Marinelli is a Senior Economist at the Education Division of the Inter-American Development Bank. Samuel Berlinski is Principal Economist at the Research Department of the Inter-American Development Bank. Matias Busso is a Principal Economist at the Research Department of the Inter-American Development Bank (email: mbusso@iadb.org). The authors declare that we have no relevant or material financial interests related to the research described in this paper. We would like to thank two anonymous referees for their detailed comments and suggestions. We thank Fundacion Luker and the Secretaria de Educacion Publica of Manizales for their support in implementing the intervention. In particular, Santiago Isaza, Maria Camila Arango and Gloria de los Rios. The pedagogical material was developed and revised over time by Alejandra Mielke, Eira Cotto, Angela Marquez and Mauricio Duque. We thank Felipe Barrera-Osorio, Jessica Gagate, Michele Giannola, Norbert Schady, and participants at various seminars for their useful comments. Anna Koh, Julian Martinez-Correa, and Juanita Camacho provided excellent research assistance. The experiments in this paper have been registered in the AEA RCT Registry # [AEARCTR-0005110](#). The data used in this article are available online: Alvarez-Marinelli, H., S. Berlinski, and M. Busso. Replication Data for: Remedial Education: Evidence from a Sequence of Experiments in Colombia. Harvard Dataverse, <https://doi.org/10.7910/DVN/HWLGXB> All appendices can be found in the online appendix to the paper, available at <http://jhr.uwpress.org/>. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent. All errors and omissions are our own.

I Introduction

Millions of children in the developing world are failing to obtain basic literacy skills during the early years of schooling (Filmer et al. 2018). In the long run, we need to address the failures of current classroom strategies to cater to the learning of children with heterogeneous needs (Bruns and Luque 2014). In the short run, there is a moral and economic imperative to find ways of addressing the needs of those who have fallen behind.¹

What is the most cost-effective strategy to educate the many children in developing countries who are failing to learn in traditional classroom settings? Targeted remedial education is perhaps the most obvious choice. However, we need to define what we are going to teach (i.e., curriculum), how to teach it (i.e., pedagogical approach), and what implementation strategy (e.g., ad hoc tutors or technology) can be scaled up in settings with relatively low teaching capabilities. In this paper, we use an evidence-based approach to design and evaluate experimentally a school-based remediation program for Spanish literacy in Colombia over several cohorts of students.

A growing literature in economics has recently focused on studying how changes in inputs affect literacy skills of school-age children. Three important insights from these large, general population studies have emerged. First, content targeted at the right level can address learning difficulties (Muralidharan, Singh and Ganimian 2019) and (Banerjee et al. 2010). Second, phonics-based methods for teaching reading (Machin, McNally and Viarengo 2018) and (Hirata and e Oliveira 2019) are particularly well suited to address the problems of those who struggle with the acquisition of literacy skills. Third, highly structured materials (Machin and McNally 2008) and (Chakera, Haffner and Harrop 2020) provide a means of ensuring fidelity of implementation and scalability, particularly in situations in which training many teachers is difficult. Our intervention includes these three elements.

The program we designed and evaluated identified struggling readers and offered them re-

¹Literacy skills matter for human capital accumulation (Zhang et al. 2014), health (Sentell and Halpin 2006), and political participation (Benavot 1996), and such skills are highly valued in the labor market (Hanushek et al. 2015).

medial literacy tutorial sessions. 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 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 both teacher-guided and independent student practice. The sessions used a curriculum 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, and reading fluency of sentences and paragraphs.

There are many important reasons to focus on reading fluency. First, reading fluency is a good indicator of reading proficiency because it is associated with comprehension in novice readers (Fuchs et al. 2001). Second, children who achieve reading fluency in early grades also tend to perform well in broader high-stakes, statewide assessments (Good, Simmons and Kameenui 2001). Moreover, 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.

Our intervention was implemented in most public schools of the Municipality of Manizales, Colombia, in three consecutive cohorts of third-grade students. For each cohort, before randomization took place, we administered an initial literacy test to identify students who were struggling to read and who were thus deemed eligible to participate in the experiments. Half of the schools were randomized into treatment and half into control groups. Tutors were randomly allocated to treatment schools, and students were randomly allocated to tutorial groups. Our research design compares eligible students in treated schools with similar, eligible students in control schools.

We report four sets of results based on the experiment. First, the intervention was effective at improving literacy skills. These gains are persistent in time and fairly homogeneous.

We find that immediately after the experiment finished (at the end of the third grade) the overall literacy scores of eligible students in treated schools improved by 0.270 of a standard deviation compared to the scores of eligible students in control schools. By the end of fourth grade, there is a reduction of forty percent on the estimated treatment magnitude for the overall literacy scores. We estimate constant quantile treatment effects for most outcomes of interest.

Second, the intervention seemed to have allowed students to better learn other subjects. We administered a standardized math test, and we find that treated children performed better in addition problems. The gains range between 0.081 and 0.104 of a standard deviation. We also find positive but not statistically significant effects on subtraction problems. These results are consistent with literacy skills fostering the acquisition of other forms of human capital.

Third, the effectiveness of the intervention increased over time. The gains by the end of the third grade on the aggregate literacy scores increased as the program went on, growing from 0.138 in cohort 1, to 0.222 in cohort 2, to 0.525 in cohort 3. These results can be explained by deliberate refinements of the program. Feedback from each cohort was used to improve the intervention effectiveness in the next wave of the intervention. 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. Some of the additional gains might also be attributed to the fine-tuning of the material, but this is difficult to quantify.

Finally, we show that the intervention is cost effective. Our intervention achieved a learning gain of 0.30 of a standard deviation per 100 dollars. As a benchmark, third-grade students in the control group increased learning by 0.18 of a standard deviation per 100 dollars spent. The program is also cost effective when costs are measured in terms of time. Students in third grade gained about 0.12 of a standard deviation per 100 hours of class, while students

in our tutorials gained a total of 0.18 of a standard deviation per 100 hours.

Our paper has three main contributions. First, we show that an evidence-based, targeted, structured pedagogical approach implemented in small-group tutorials in developing countries can have significant effects on the learning trajectories of struggling children. In this sense, our paper is close to [Banerjee et al. \(2007\)](#); their paper finds positive effects of a remedying tutorial intervention targeted to low-achieving students. However, their study differs from ours in several important aspects. They report results of a year-long program that provided struggling students (as determined by the schools) with two hours a day of tutoring support. Our intervention is less intensive and therefore easier to scale up. There are methodological differences, too. [Banerjee et al. \(2007\)](#) do not have information on ex ante eligibility for children in the control schools; we have this information, and we can therefore identify the treatment effect of the intervention on the group of students who are at risk of illiteracy. This is arguably an important policy-relevant parameter of interest.

Our paper is also close to [Muralidharan, Singh and Ganimian \(2019\)](#), whose work finds positive effects of a computer-based personalized instruction program that was delivered after school six days a week for 90 minutes each day. The personalized instruction program anchored content to the level of knowledge of each individual student and therefore allowed students to be taught at the right level. Although promoting reading fluency is less amenable to the use of technology, our paper complements this study by showing that schools in low-capacity settings can teach at the right level by delivering targeted, structured content in small-group tutorials.

Our approach is similar to that pursued by [Johnson et al. \(2019\)](#), who look at the impact of small group tuition for 5-year-old pupils in English schools. The design of the curricular approach in their work revolves around a meta-analysis of successful interventions in literacy. Teaching assistants already employed at the school were trained to deliver the program for 15 minutes of teaching four times a week over 20 weeks. Their randomized evaluation of 50 schools finds a short-term impact on children's reading scores that tends to fade out

over time. The crucial difference between their study and ours is that the program used in [Johnson et al. \(2019\)](#) is not remedial; it was used for all 5-year-old students.

A second contribution of our paper is to provide new evidence that reading remediation using a phonics approach works at scale. There are some small-scale studies in controlled settings in developed countries that look at reading remediation early in elementary school ([Slavin et al. 2009](#)). It is an open question as to whether remediation works at scale in less controlled environments, particularly in developing countries with less qualified teachers ([Kerwin and Thornton 2019](#)). We show in three consecutive experiments that these types of programs are indeed effective.

The third contribution, a methodological one, is to showcase the benefits of sequential experimentation for policy analysis and design. By repeating an experiment under different implementation protocols, we increase our confidence on the effectiveness of the intervention's main ingredient (a small-group remediation program with a structured, phonics approach). Furthermore, by using variation across cohorts in implementation protocols we are able to learn about what secondary aspects of the intervention may improve its effectiveness. We provide a useful case study for economists, who increasingly interact with policy makers in the design of public policies ([Duflo 2017](#)).

The rest of the paper is organized as follows: Section [II](#) describes the intervention and the setting in which it took place. Section [III](#) presents the experiment and the data. Section [IV](#) presents the main results of the paper and discusses implications. Section [V](#) presents results that explain the increased effectiveness of the intervention over time. Section [VI](#) provides calculations of the cost effectiveness of the intervention. Section [VII](#) concludes.

II Intervention

II.A Setting

The sequence of experiments took place among third-grade students of public elementary schools in the Municipality of Manizales in Colombia over three consecutive years (2015-2017). Manizales is a mid-size city. Approximately 13.8 percent of residents have incomes below the poverty line, and 6.9 percent of the municipality’s residents live in rural areas. About 97 percent of the children participating in the study can be considered to be disadvantaged.² More than 18,000 children were enrolled in the first five grades of the public elementary school system.³

Public schools operate in either six- or eight-hour shifts for 165 days a year. During regular school hours students receive instruction according to the primary school curriculum which includes four main academic subjects: Spanish, mathematics, natural sciences, and social sciences. In addition to learning these main subjects, students also have instruction in other subject, including art, physical education, and technology. In early grades academic subjects are all taught by the same teacher. Although there are national guidelines regarding what children should achieve, schools and teachers are at the liberty of choosing pedagogical approaches and classroom strategies (MEN 2016).

The municipality’s third graders scored slightly above the national mean 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 was to create a remedial program, based on the phonics approach, to improve the reading fluency of struggling

²In our sample, 97 percent of students fall in levels zero to three of the socioeconomic status classification scale used in Colombia to target social programs.

³Schooling in Colombia is compulsory from kindergarten to ninth grade. 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.

third-grade students. In general, regular class teaching of literacy in Colombia incorporates a hodgepodge of approaches to teach children to read. Teaching combines a “whole language” method with some syllabic components, rather than taking a phonics approach.⁴

II.B Reading Remediation in Small Tutorial Groups

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 and 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 children need certain foundational skills. There is a strong consensus from research on reading instruction (see, for example, Foorman and Torgesen (2001) and NAEP (2000)) that these necessary skills include: 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. Moreover, there is also consensus that these skills should develop in the early years of school. Good, Simmons and Kameenui (2001) propose a timeline for the development of these skills: phonological awareness during kindergarten, decoding and acquiring the alphabetical principle in first grade, and gaining accuracy and fluency when reading in second and third grades.

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, Simmons and Smith 1998, Stanovich 1986). This is because children at risk of reading failure acquire reading skills more slowly than other children. Thus, students lacking foundational

⁴There is a large literature that debates the benefits of using the “whole language” versus phonics approach for early literacy (Soler 2016). The pendulum has now swung in favor of phonics (NAEP 2000).

reading skills by third grade will not be able to read fluently, and it is difficult for them to catch up to their peers. 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).

Our intervention was designed to satisfy those requirements. The intervention followed structured class materials. 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., the 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 delivery of the intervention was in the form of small-group tutorials (no more than six students). There is evidence that one of the most practical methods for increasing instructional intensity for a small number of at-risk students is to provide small-group instruction. This also allows for the tutor to anchor the lessons to the level of each individual student. 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 ([Elbaum et al. 2000](#)), an interesting finding emerging from these analyses is that one-to-one interventions in reading are not necessarily more effective than small-group interventions.

Tutors led tutorials in a designated school space. There is evidence in education (see [Elbaum](#)

et al. 2000) that many successful interventions can be delivered by trained individuals rather than reading specialists. Each tutor oversaw an average of five tutorial groups each year. Fifteen tutors were hired each year of the intervention.⁵ The recruitment process was simple. After a call for applications distributed through local schools and universities, the local NGO received a large number of resumes. Candidates were then selected based on those resumes and through simple interviews. At the beginning of each year of the intervention, prospective tutors received an eight-hour training session. These sessions were delivered by trainers who were involved in the design of the material, by other local reading specialists, and, in later rounds, by former tutors. 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. The feedback received in these sessions was used to fine-tune the material used in the tutorials in later cohorts.

Most tutors were students pursuing a teaching career, or recent graduates who had not yet entered the formal career (which is heavily regulated). The median tutor was 26 years old, and 97 percent of them were women. They were paid, by the local NGO, an hourly wage equivalent to that obtained by a teacher who had just entered a teaching career with a similar level of education. In all three cohorts, there was a large excess supply of suitable applicants willing to work as tutors. Thus, conditional on the budget constraint, it would be easy to scale up the intervention even further to other cities in the country.

The tutorials happened during regular school hours. Struggling readers were taken out of the classroom three times a week for 40 minutes during these regular school hours. The intervention was implemented during the second half of each academic year (starting right after the June mid-year break). In the case of the first experimental 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). While these tutorials took place, students in the control group continued receiving instruction as usual. The decision about when to

⁵During 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.

take the students out of the classroom was based on logistical constraints and teachers were not consulted. We have no information on what classes children may have skipped to attend tutorials.

III Research Design

III.A Measures

We measure language development using the Early Grade Reading Assessment (EGRA), which was designed by [RTI-International \(2009\)](#) under the auspices of the U.S. Agency for International Development (USAID) and the World Bank. This open-source assessment tool has been applied in more than 65 countries for countrywide assessments and program evaluations ([Dubeck 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 most subtasks; if a child is unable to finish the subtask in that time, she moves to the next subtask.⁶

We collected information on the following EGRA subtasks: (i) knowledge of letter sounds (which requires students to sound the letter), (ii) reading of non-words (which requires students to string letter sounds in words that do not have any meaning), (iii) fluency of oral reading (which requires students to read a paragraph aloud either by recognizing words by sight or by reading its phonemes), and (iv) reading comprehension (which asks students to respond questions about the paragraph they read for the previous subtask).⁷ We also used

⁶For most subtasks, the items within it are, a priori, of equal difficulty. The comprehension subtask is not timed.

⁷The Ministry of Education of Colombia established a set of literacy skills that students should be acquiring as they progress through their basic education. These include many of the literacy skills tested in EGRA, for instance: parsing sounds present in different words (first grade), identifying the repetition of sounds at the end of sentences (first grade), reading simple words (first grade), associating a letter or groups of letters to a sound when reading out loud (first grade), identifying syllables that form a word and how they relate to the location of written accents (second grade), fluently reading out loud simple texts written by the student themselves or by others (third grade), comprehending the content and the structure of a text,

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, in one-on-one sessions 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 Online 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.⁸ At the beginning of the academic year we collected information about students in third grade. 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 third grade and at the beginning/middle and end of fourth grade.

III.B Sample

The study was designed to be implemented in the universe of schools in the municipality. Ninety-four schools participated in the experiment in the first cohort. In the second and third cohorts, we eliminated from the sample the schools at the extremes of the size distribution.⁹ We used the information collected at the beginning of the school year (before randomization took place) from the universe of schools in the municipality to determine which students were considered eligible to participate in the tutorials.

Eligibility was determined based on a reading baseline test score. In the case of the first and making inferential or critical claims (third grade) (MEN 2016). For this reason, non-treated students should be able to develop many of the skills tested in EGRA through normal class instruction.

⁸All data and code used for the analyses in this paper are available via the Harvard Dataverse (Alvarez-Marínelli, Berlinski and Busso 2021).

⁹We started with a universe of 94 schools in 2015. During the first year we found that that groups that were too small or too large added too much logistical complexity to the intervention. The experimental sample was 84 schools (in cohort 2) and 80 schools (in cohort 3). In the Online Data Appendix (section II.A) we present the average observable characteristics for the sets of school that participated in the experiment for each of the three cohorts. We are not able to reject any of the null hypotheses of equality in means in those variables.

cohort, student eligibility to participate in the tutorial sessions was determined in two steps. First, we constructed an equally weighted composite index of the following EGRA subtasks: reading of non-words, fluency of oral reading, and reading comprehension. Using this index we determined the number of students who scored below the 25th percentile in each school. This number was then used to determine the number of tutorial sessions that would be offered in each school. Second, we selected students who were closest to the 25th percentile threshold to complete tutorial groups to maximum capacity. Notice that this procedure was followed in *all* schools, before they were randomized to treatment.

In the case of the second and third cohorts we used a slightly different eligibility criterion. 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.¹⁰ Again, this eligibility criterion was applied before randomization to all schools.¹¹

Figure 2 depicts the distribution of baseline scores and the threshold for eligibility in the universe of children.¹² The Online Data Appendix presents information on the sample sizes and the response rates of each cohort.

III.C Randomization and Experimental Validity

Randomization of the treatment was done at the school level. First, we sorted schools based on how many children were eligible for treatment. We created blocks of two schools and then, within these blocks, we randomized schools to treatment and control. We repeated this procedure 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 eligible students in control schools carried on with their usual classroom learning experiences.

¹⁰We did not attempt to fill tutorial groups to full capacity by adding students that did not comply with the 60 words cut-off

¹¹Eligibility was strictly enforced. Principals and teachers could not add students to tutorials; nor could they prevent eligible students from participating in them.

¹²Using a two-sample Kolmogorov-Smirnov test, we cannot reject the null of equality of distributions of the baseline scores between treated and control students in any of the three cohorts.

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

$$W_{is} = \theta T_s + \mu_{strata} + \epsilon_{is} \quad (1)$$

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. 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. Demographic characteristics, as well as reading and math scores at baseline, were not statistically different for eligible students in treatment and control schools in any of the three cohorts.¹³ Most of these students attended medium-size, urban or periurban schools during the morning shift. The schools were low income.¹⁴

Once students were deemed eligible, they were assigned to tutorial groups in their school. In the case of the first cohort, when there were more than six eligible children, tutors and schools organized the compositions of the tutorials. We modified the tutorial assignment in the second and third cohorts: In schools with more than six eligible students, students were assigned randomly to equally sized tutorials. In each school there was only one tutor. Tutors were randomly assigned to schools in each year of the experiment. Because tutorial

¹³The reading comprehension subtask included five questions in cohorts 1 and 2, and eight questions in cohort 3. This explains the larger value for this latter cohort.

¹⁴We measure the socioeconomic status using a categorical variable (*estrato*) that takes six values: 1 (very low income), 2 (low income), 3 (medium-low income), 4 (middle class), 5 (upper-middle class), and 6 (upper class). These values are assigned to households as part of the System of Identification of Potential Beneficiaries of Social Programs (SISBEN).

assignment was done before schools were randomized to treatment, we know the hypothetical assignment for eligible students in control schools as well.¹⁵ Panel B of Table 1 describes the tutorial groups. The average tutorial size was the same in both treatment arms in all cohorts. However, because we filled tutorial groups to full capacity in cohort 1, but not in cohorts 2 and 3, the groups were larger in the first cohort. This difference in size did not affect the heterogeneity of the tutorial group across cohorts, as captured by the standard deviation of the literacy baseline score of the students in the tutorial. It did affect the average scores of students. These scores are higher for the first cohort – an issue to which we return in Section V.

Once the tutorial sessions started, compliance to treatment was high. Panel C of Table 1 describes the compliance with treatment, as shown by 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.

After the tutorials finished, 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.¹⁶ Panel D 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. We find 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 fourth grade is 54-60 percent. This was due to a logistical problem with the data collection that prevented administering

¹⁵In the case of the first cohort, the assignment to tutorials was done by the schools. We assigned students to tutorial groups in control schools at random.

¹⁶Very few students dropped out of school.

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. It is reassuring to find that attrition was not correlated with baseline reading ability or with the treatment status (as we show in the Online Data Appendix). All in all, we think that attrition is not a major concern or a threat to identification of treatment effects.¹⁷

IV 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.

IV.A Empirical Strategy

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

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

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 ($h = 3$).¹⁸ μ_c are strata fixed effects 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).

¹⁷Despite this, in Section IV.C we assess the effect of attrition on the treatment effect estimates.

¹⁸Some students repeat, and therefore they are observed twice in the third grade. Thus, the model also includes year dummies.

T_{sc} is 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 that takes the 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.¹⁹

IV.B Main Results

We plot in Figure 3 the raw count of correct answers in the literacy test, 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 third grade, and 133 items by the end of fourth grade.²⁰ By contrast, students in treated schools correctly answered 124 (end of third grade) and 140 (end of fourth grade) items. The figure suggests that the treatment group experienced positive gains from this intervention, and that the gains persisted over time.^{21,22}

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

¹⁹Standard errors are very similar when we estimate cluster them at the school-cohort level (i.e., the unit of randomization, given that we randomized schools to treatment in each cohort of the experiment).

²⁰The fact that students in the control group continue to develop literacy skills over time, as measured by EGRA, suggests that normal classroom instruction by itself is effective. As a corollary, this implies that our outcome measures are “fair” to the control group (in the sense that they do not target skills solely developed in the small-group tutorials).

²¹The growth rate is much higher between the beginning and end of the third grade than in the fourth grade because subtasks become relatively easier once a minimum of reading fluency is achieved.

²²We estimated the effect of the intervention on the achievement gap between ineligible and eligible students on the raw literacy score (defined as the number of correct answers in all EGRA’s subtasks). The gap is the same for treated and control students at the beginning of third grade (the ineligible scored 48.5 more correct answers than the eligible). The gap closes for all students during third grade (reaching 34.9 in the control group at the end of third grade). However, consistent with a positive treatment effect, it closes more for those students in treated schools (reaching 26.1). It then remains constant during the summer break and increases a slightly during fourth grade, both for treated (reaching 28.7) and control (reaching 36.4) students.

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 0.349 of a standard deviation (or four-letter sounds).²³ 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 correctly read an average of 27 non-words. The treatment effect is 0.073 of a standard deviation, which translates into a gain of less than one 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 find that treated children’s reading scores were 0.157 of a standard deviation higher (representing a gain of more than two 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 ([Whitehurst and Lonigan 1998](#), [Foorman and Torgesen 2001](#)). One of the skills involves the ability to manipulate letters, sounds and phonemes. The other includes vocabulary and conceptual knowledge. Both are key to master the skill of 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 aggregate 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 0.270 standard deviations. We take this effect to be quite large considering that the gain

²³Note 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 fourth grade for the first cohort.

during third grade of the average student in schools of the control group is 0.400 standard deviations.²⁴

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](#), [Chetty et al. 2011](#)). Because reading is a skill that 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 that 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.

In sum, using a literacy index that adds the number of correct responses in each subtask we find a gain of 0.270, 0.264, and 0.152 of a standard deviation at the end of third grade, beginning of fourth grade, and end of fourth-grade, respectively. All estimates 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 effect persisted one year after the intervention finished.

IV.C Robustness

We assess the robustness of our results along three dimensions. In the interest of saving space, Table 3 shows the different treatment-effect estimates only at the end of third grade.²⁵

²⁴The average number of correct responses, among all students attending schools in the control group, to the literacy subtasks was 126 at the beginning of third grade (with a standard deviation of 35.5) and 140 at the end of the third grade. This results in a gain of $(140 - 126)/35.5 = 0.394$ of a standard deviation.

²⁵The full set of results for outcomes measured at the beginning and end of fourth grade can be found in the Online Appendix I.A.

First, we show that our main results are robust to several changes in the model’s specification (Panel A). We start by adding different sets of pre-treatment control variables to equation (2). Neither including the corresponding baseline test scores nor including individual and school controls affects the results. This is not surprising given that we have shown that these variables were balanced before treatment. Further, we condition on school fixed effects by exploiting the fact that a new randomization was implemented in every cohort. Reassuringly, the intention-to-treat estimates remain unchanged.

Second, we show that students’ attrition has very little effect on our main results (Panel B). As we discussed in Section III, we found overall low attrition rates that were similar for eligible students in treated and control schools. We estimate lower and upper bounds for our main treatment effects following Lee (2009). We find that the lower bound for the treatment effect remains positive for all outcomes (except reading comprehension), and that the literacy score remains statistically significant at standard levels at the end of third grade and throughout fourth grade (not shown in the table).

Third, Table 2 focuses on four outcomes measured at three different time horizons, which could raise some concerns about multiple hypotheses testing. In our research design, these concerns are mitigated for two reasons. First, we pre-specified our main outcome variables. From the outset, before the intervention started, our main outcome of interest was “fluency of reading”. Secondary outcomes included other off-the-shelf subtasks that were part of the EGRA instrument. We aggregated all subtasks into one literacy score. Second, rather than estimating different equations for different measures at different time horizons, we estimated treatment effects at all time horizons *jointly* in one parsimonious model. Despite these advantages of our research design, we undertake a robustness check by controlling for the family-wise error rate following Westfall and Young (1993) for the *full* set of outcomes and time horizons. (Panel C shows the adjusted p-values for end of third grade only.) The treatment effects of the intervention at the end of third grade, as well as the effects on literacy scores at the beginning and end of fourth grade (not shown in the table), remain

statistically significant at standard levels.

IV.D Indirect Effects of the Intervention

The tutorial required students to be taken out of the classroom. This could have negatively affected their classroom learning because they received 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, McNally and Viarengo \(2018\)](#), for example, find spillovers to mathematics from interventions that successfully change reading skills.

Table 4 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 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,²⁶ we do not find that the intervention affects the probability of repeating the grade.

The intervention can also have an indirect effect on the classmates of eligible students (i.e., the students that performed better in the baseline tests and were ineligible to participate in the small-group tutorials). In a companion paper, [Berlinski, Busso and Giannola \(2021\)](#) quantify these spillover effects and find that the literacy scores of ineligible children in treated schools increased; the increase among ineligible students was a third of the size of the increase among eligible students who received the treatment. The study also reported no spillover effects on the mathematics scores of ineligible children. In addition to estimating

²⁶There is no evidence of grade progression being guaranteed in Manizales. On average 3 percent of students, or about 1 in 30 students, are retained during the first five grades. Not surprisingly, retention rates are higher in the sample of eligible students.

these reduced-form effects, the paper exploited the randomization to identify peer effects in the classroom from low-achieving to high-achieving students. Using a linear-in-means model of peer effects, the authors found that a one-standard-deviation increase in peers' contemporaneous achievement increased individual test scores by 0.679 of a standard deviation.²⁷

V Sequential Experimental Results

The findings summarized in the previous section are the result of a sequence of experiments. Repeating the experiments with different cohorts, keeping constant the main ingredient of the intervention (a small-group remediation program with a structured, phonics approach) but under different treatment protocols has two advantages. First, it ensures a higher degree of replicability and external validity than a one-shot experiment. Repeated success or failure provides higher confidence on the effectiveness of the main ingredient, and suggests that the results are not driven by exceptionally favorable or unfavorable circumstances. Second, it allows us to use feedback from each cohort to vary some secondary implementation aspects to increase the effectiveness in the subsequent waves of the intervention. We can combine across-cohort and within-cohort variability to understand what elements of the intervention (mechanisms) contributed to its success or failure.²⁸

We estimate the treatment effects for each cohort, outcome, and time horizon combination. A total of 43 parameters are reported in Appendix Table A.4. 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.²⁹ There is an upward trend over time in the treatment effect of the intervention. The gains by the end of the third grade on the aggregate literacy score grow from 0.138 in cohort 1, to 0.222 in cohort 2, and

²⁷Identification of the structural peer effect parameter required ruling out alternative explanations coming, for instance, from a reduction in class size.

²⁸The alternative to this is to vary the secondary aspects of the implementation by randomizing schools to multiple treatment arms. However, this approach would require ex ante knowledge of which secondary implementation aspects are relevant. It would also require a large number of schools to provide enough statistical power to identify statistically significant differences across treatment arms.

²⁹Results are similar when the estimates are weighted by the inverse of the standard error.

0.525 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 0.056, to 0.180 from cohorts 1 to 3.

Our analysis of results from the first cohort showed positive treatment effects on knowledge of letter sounds only (see Appendix Table A.4). These results were to some extent disappointing because reading fluency, our main target outcome, did not improve. As a consequence, the research and implementation teams identified several areas where improvements in the intervention could be made for the subsequent rounds. The aims of these changes were to increase intensity and to improve the cognitive support of the intervention. 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. We reviewed the pedagogical material, replacing some of the vocabulary development tasks with 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 all, we introduced four changes after the first cohort. We increased the dosage. We modified the targeting. We fine-tuned the material. And we modified the assignment of students to tutorial groups, which could potentially lead to changes in the composition of the tutorial groups. There was a fifth change for cohorts 2 and 3: In the wake of the experiment with the first cohort, some tutors for subsequent cohorts had previous experience, the result of having delivered the intervention previously. Next, we analyze how these five factors may have contributed to the increased impact of the intervention over time. We cannot, however, isolate the contribution that each of these factors had to the increased effectiveness of the program because many of them occurred concurrently as the program, informed by the initial experimental results, evolved over time. Instead, we exploit the institutional knowledge that comes from working closely with the policy makers to provide some additional results and back-of-the-envelope calculations to obtain a sign and an upper bound for their contributions.

V.A 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.

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

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

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 $h = 1, 2, 3$). Like equation (2) this equation includes strata-cohort and time-horizon fixed effects.

Actual tutorial attendance might not be orthogonal to ϵ_{isch} , for instance, because initially lower-achieving, eligible students might skip school-days more often. Thus, adopting an approach similar that that used by [Muralidharan, Singh and Ganimian \(2019\)](#), we instrument D_{isc} with the randomized treatment variable T_{sc} (interacted with the time-horizon indicator variables), which is likely orthogonal to ϵ_{isch} . Equation (3) estimates a dose-response function using the within- and between-cohort variation in exposure. We believe that the exclusion restriction holds in our setting for two reasons. First, treatment assignment was randomized.

Thus, any violation of the exclusion restriction requires that the randomization itself affected students' literacy skills by changing the behavior of teachers or students. We think this is unlikely: In each cohort randomization was done at the school level by the research team; it was not publicly announced. Teachers and students in control schools were not aware that they had been randomized to control status. Second, some of the variation in dosage stems from modifications to the intervention design (i.e., the increase in the intensity of the treatment and the offering of make-up sessions). These modifications were unlikely to affect literacy skills other than through the availability of more tutorial sessions.

In Table 5 we present the instrumental variables results.³⁰ There is a positive dosage effect. At the end of third grade, students who attended one additional session performed 0.007 of a standard deviation better in the literacy test than students in the control group.³¹ These results decay slightly in fourth grade. 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 translates into a gain of $0.007 \times 17 = 0.119$, which is similar to the difference in the estimated treatment effect between cohorts 1 and 2 (i.e., $0.222 - 0.138 = 0.084$), and about a third of the increase between cohorts 1 and 3 (i.e., $0.525 - 0.138 = 0.387$).

V.B 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-word reading subtask for a longer (132-word) text.³² 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 6 shows the average differences in performance (not standardized) on these outcomes. We find that students in cohorts 2 and 3 had lower levels

³⁰Our first stage is very strong. The F-Stat is higher than 1828 in all models.

³¹Results are similar when estimated using OLS.

³²In addition, for cost reasons, we eliminated from the experimental sample schools with only one eligible student.

of skills than those eligible in cohort 1, and that students in the third cohort performed better at baseline 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 among the three cohorts. Figure 6 investigates this by estimating quantile treatment effects following [Firpo \(2007\)](#).³³ 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 the last two cohorts of the experiment.^{34,35}

V.C Tutorial Composition

A third change addressed the composition of the students who attended tutorials. In all three cohorts the size of the tutorials were determined by the researchers (not by the school). The composition of the tutorials, however, could have changed: In the first cohort of the experiment we allowed the NGO to assign students to tutorials based on logistical considerations. 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 3.³⁶

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 7 shows the treatment effects for two groups and

³³For an application of the estimation of quantile treatment effects see, for instance, [Bitler, Gelbach and Hoynes \(2017\)](#).

³⁴Results, 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.

³⁵In Appendix Table A.5 (Panel A) we explore treatment-effect heterogeneity by students' sex. We find that for all literacy outcomes the estimated treatment effects for girls are larger than those for boys. However, we cannot reject equality of the estimated coefficients.

³⁶Of 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 explain the gains we observe over time.

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.³⁷ 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 heterogeneous 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.³⁸ 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.

V.D 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 7 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 0.145 of a standard deviation more in the overall literacy score.³⁹ However, this difference is not statistically significant. These

³⁷Students are classified according to the *observed* number of students in the tutorial.

³⁸In Appendix Table A.5 (Panel B) we show that the point estimate of the treatment effect is also larger for those students that come from more homogeneous *classrooms*. However, we are only able to reject the null hypotheses of equal treatment effects at normal confidence levels for “reading of non-words” and “reading comprehension.”

³⁹Because the allocation of tutors to schools was randomized in each cohort these results are not driven by differential tutor dropout from worse-performing school.

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 to different activities between more and less experienced tutors.

V.E 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.⁴⁰ In addition to these changes, in the third cohort, 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. This is, of course, only speculative. It could be that other contemporaneous

⁴⁰Students 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.

factors such as the work of the institutional participants changed as the program evolved. We do not think this is the case. The municipal government and the NGO, for instance, had little involvement in the day-to-day activities beyond the work of the tutors, or the decisions regarding format of the tutorials or the materials. The schools had also little room to adjust their behavior. The tutorials happened in the second semester of the school year when most decisions regarding classroom assignment of teachers and students had already been made.⁴¹

VI 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 of a 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 0.270σ 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.⁴² 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.⁴³ 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 can teach more children per day (by offering more sessions).

A second policy-relevant indicator of cost effectiveness would be to compare learning gains

⁴¹Other contextual factors, such as the political climate, the weather, or any other municipal-level event would have equally affected both treated and control schools, and therefore could not explain the increasing effectiveness of the program.

⁴²See Appendix Table [A.6](#)

⁴³According to the World Development Indicators, GDP 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 $89/5871 \times 100 = 1.5$

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 0.30σ per 100 dollars. However, as noted by [Muralidharan, Singh and Ganimian \(2019\)](#), while spending in education can increase unboundedly over time, students are in school only for a set number of hours a day. For this reason, it is also relevant to evaluate the effectiveness of the program in terms of its time costs. Again, students in third grade gained about 0.12σ per 100 hours of class, while students in our tutorials gained a total of 0.18σ per 100 hours.⁴⁴

VII Conclusion

In countries where many students are reading below grade level 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. 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 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 persisted. 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

⁴⁴Recall from Section IV that students in our program gained 0.270σ , and that during third grade, students 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 yields a per 100 dollar effect of $(100 \times 0.4)/(665/3) = 0.18$ for students in the control group and $(100 \times 0.270/89 = 0.30)$ for treated students. Similarly, this yields a per 100 hour effect of $(100 \times 0.4)/(1000/3) = 0.12$ for students in the control group and $(100 \times (0.4 + 0.27))/(1000/3 + 32) = 0.18$ for the students in the tutorials; this assumes, conservatively, that 32 hours were additional hours of literacy.

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 make several changes to address potential factors that likely explained our limited initial success. 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 (such as changing the way reading is taught in earlier grades of school so that fewer children reach the third grade still struggling to read). Taking similar steps earlier could be even more cost effective.

References

- Alvarez-Marinelli, H., Berlinski, S. and Busso, M. (2021), ‘Replication data for: Remedial education: Evidence from a sequence of experiments in Colombia.’, *Harvard Dataverse* .
- Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R. and Khemani, S. (2010), ‘Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India’, *American Economic Journal: Economic Policy* **2**(1), 1–30.
- 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.
- 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.
- Berlinski, S., Busso, M. and Giannola, M. (2021), ‘Helping struggling students and benefiting all: Peer effects in primary education’, *mimeo* .
- 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.
- Bruns, B. and Luque, J. (2014), *Great teachers: How to raise student learning in Latin America and the Caribbean*, The World Bank.
- Chakera, S., Haffner, D. and Harrop, E. (2020), Structured pedagogy: For real-time equitable improvements in learning outcomes, Working Paper 10701, UNICEF.
URL: <https://www.unicef.org/esa/media/7511/file/ESA-Structured-Pedagogy-2020.pdf>
- 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.
- 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.
- Filmer, D., Langthaler, M., Stehrer, R. and Vogel, T. (2018), 'Learning to realize education's promise', *World Development Report. The World Bank* .
- Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**(1), 259–276.
- Foorman, B. 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.
- 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.
- Hamushek, 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* .
- Johnson, H., McNally, S., Rolfe, H., Ruiz-Valenzuela, J., Savage, R., Vousden, J. and Wood, C. (2019), 'Teaching assistants, computers and classroom management', *Labour Economics* **58**, 21–36.
- Kerwin, J. and Thornton, R. (2019), 'Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures', *unpublished* .

- Lee, D. S. (2009), ‘Training, wages, and sample selection: Estimating sharp bounds on treatment effects’, *The Review of Economic Studies* **76**(3), 1071–1102.
- 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.
- MEN (2016), *Derechos Básicos de Aprendizaje: Lenguaje*, Ministerio de Educación Nacional.
- 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 (2019), *Education at a Glance 2019*.
URL: <https://www.oecd-ilibrary.org/content/publication/f8d7880d-en>
- 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. and Shaywitz, B. (1996), ‘Unlocking learning disabilities: The neurological basis’, *Learning disabilities: Lifelong issues* pp. 255–260.
- 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).
- Soler, J. (2016), ‘The politics of the teaching of reading’, *Prospects* **46**(3), 423–433.
- Stanovich, K. (1986), ‘Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy’, *Reading Research Quarterly* **22**, 360–407.
- Westfall, P. H. and Young, S. S. (1993), *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Vol. 279, John Wiley and Sons.
- 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.

Figures

Figure 1: Timeline (for Each Cohort)

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 who performed approximately 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 who 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 six students) with students immediately next to the 25th percentile threshold. In cohorts 2 and 3 the eligibility criterion was strict (represented by a solid line).

Figure 3: Effect of the Intervention on Literacy Score

Note: The solid black line shows the number of correct answers by eligible students in schools randomized to treatment. The dashed gray line shows the number of correct answers by students in schools randomized to control. Confidence intervals at the 95% level. “Beg G3” refers to the measure taken in March (baseline) of third grade, “End G3” refers to the measure taken by the end of third grade, after treatment. “Beg./Mid. G4” refers to the first measure taken in fourth grade, “End G4” refers to the measure taken by the end of fourth grade. The vertical dotted line marks the approximate time of treatment.

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.4 for the individual estimated treatment effects.

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 II 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.

Figure 6: Quantile Treatment Effects

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

Tables

Table 1: Balance, Attrition, and Compliance

		Cohort 1			Cohort 2			Cohort 3		
		T	C	p-value	T	C	p-value	T	C	p-value
Panel A:	Avg. student's pre-treatment characteristics									
	Age	8.82	8.83	0.71	8.49	8.37	0.09	8.40	8.47	0.31
	Proportion female	0.50	0.48	0.74	0.47	0.50	0.42	0.50	0.52	0.76
	Scores:									
	Fluency of oral reading	49.84	50.73	0.14	42.17	42.12	0.98	44.17	43.42	0.51
	Knowledge of letter sounds	14.82	13.64	0.20	12.41	13.46	0.18	19.51	19.00	0.69
	Reading of non-words	28.83	28.33	0.54	24.22	24.88	0.41	26.62	26.83	0.84
	Reading comprehension	3.07	3.10	0.55	2.10	1.99	0.20	5.33	5.05	0.26
	Literacy score (avg. of sub-scores)	0.59	0.59	0.90	0.39	0.39	0.37	0.49	0.48	0.55
	Addition	12.18	12.23	0.60	10.15	9.97	0.57	11.78	11.89	0.67
	Subtraction	9.51	9.21	0.40	8.12	7.74	0.26	9.44	9.16	0.35
	Math score (avg. of sub-scores)	0.42	0.42	0.95	0.35	0.34	0.31	0.41	0.41	0.87
	Avg. school's pre-treatment characteristics									
	Grade 3 enrollment	47.71	47.70	0.83	41.33	42.88	0.76	45.83	43.45	0.54
	Morning	0.79	0.72	0.50	0.74	0.76	0.80	0.77	0.70	0.49
	Rural	0.25	0.20	0.53	0.12	0.17	0.57	0.15	0.17	0.74
	Avg. Socioeconomic status (1-6)	1.89	1.81	0.44	1.81	1.91	0.26	1.89	1.82	0.51
	Avg. class size	23.91	23.34	0.56	23.42	26.06	0.09	24.65	23.49	0.48
	Avg. number of classrooms in Grade 3	1.56	1.61	0.84	1.45	1.36	0.43	1.55	1.50	0.69
Panel B:	Tutorial assignment									
	Avg. tutorial size	5.92	5.91	0.63	4.82	4.79	0.67	4.99	4.91	0.30
	S.D. literacy score of students in tutorial	0.09	0.10	0.32	0.10	0.10	0.71	0.10	0.10	0.79
Panel C:	Compliance									
	Ever attended a tutorial	0.92	0.00	0.00	0.94	0.00	0.00	0.98	0.00	0.00
	Tutorial attendance (percent)	0.73	0.00	0.00	0.90	0.00	0.00	0.90	0.00	0.00
	Tutorial attendance (sessions)	26.20	0.00	0.00	43.03	0.00	0.00	42.99	0.00	0.00
Panel D:	Attrition									
	End of Grade 3	0.06	0.07	0.19	0.07	0.09	0.27	0.09	0.09	0.95
	Beg./Mid. Grade 4	0.54	0.60	0.25	0.12	0.16	0.08	0.10	0.12	0.40
	End Grade 4	0.21	0.17	0.09	0.14	0.12	0.50	0.14	0.18	0.20

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 subtask. Panel B shows tutorials' characteristics. In the case of the control group we assigned students to hypothetical tutorials (even when no tutor was assigned to them). Panel C shows which eligible students attended the tutorials and the intensity of attendance. Panel D shows the attrition rates at different time horizons of the eligible students observed at baseline. 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.

Table 2: Treatment Effects on Literacy
(standardized literacy outcomes)

	Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading com- prehension	Literacy score
	(1)	(2)	(3)	(4)	(5)
Treatment x End of Grade 3	0.349*** [0.059]	0.073 [0.050]	0.157*** [0.049]	0.001 [0.044]	0.270*** [0.056]
Treatment x Beginning of Grade 4	0.312*** [0.069]	0.105* [0.056]	0.192** [0.077]	0.150** [0.062]	0.264*** [0.075]
Treatment x End of Grade 4	0.300*** [0.055]	0.088** [0.041]	0.083 [0.053]	0.055 [0.053]	0.152*** [0.051]
Observations	4949	6362	6362	6362	6362
p-value of equal coeffs.	0.726	0.881	0.272	0.103	0.065
Baseline control mean (non-std.)	15.02	26.94	46.32	3.303	0.507
Baseline control S.D. (non-std.)	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 $H_0 : \theta_1 = \theta_2 = \theta_3$. Control mean and S.D. correspond to the number of correct answers (cols. 1-4) or the % of correct answers (col. 5). 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: Robustness
Treatment Effects at the End of Grade 3 (standardized outcomes)

	Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading com- prehension	Literacy score
	(1)	(2)	(3)	(4)	(5)
Panel A: Adding controls					
Baseline outcome	0.342*** [0.054]	0.073 [0.045]	0.168*** [0.048]	-0.002 [0.043]	0.273*** [0.049]
Individual- and school-level variables	0.336*** [0.058]	0.078 [0.047]	0.167*** [0.047]	0.006 [0.043]	0.274*** [0.054]
School fixed effects	0.448*** [0.097]	0.117 [0.086]	0.147* [0.079]	0.044 [0.075]	0.307*** [0.092]
Panel B: Lee-bounds					
Lower bound	0.328*** [0.050]	0.042 [0.044]	0.127*** [0.048]	-0.086** [0.043]	0.234*** [0.048]
Upper bound	0.371*** [0.049]	0.106** [0.050]	0.192*** [0.053]	0.034 [0.056]	0.307*** [0.054]
Panel C: Multiple Hypothesis Testing					
Treatment	0.349***	0.073	0.157*	0.001	0.270***
Adjusted p-value	0.001	0.522	0.087	0.983	0.008

Note: The table compiles different robustness checks for estimated treatment effects at the end of Grade 3. Panel A adds controls for the main specification. Individual-level controls include age, age squared, gender, socio-economic and disability status. School-level controls include class size, school number of classes in grade 3, rural and morning school-status. Panel B estimates Lee (2009) bounds. Panel C calculates Westfall-Young adjusted p-values to address for multiple hypotheses testing. Appendix Tables A.1, A.2 and A.3 present details and the same robustness for beginning and end of Grade 4. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: Treatment Effects on Other Education Outcomes

	Additions (std.)	Subtractions (std.)	Math score (std.)	Repeat Grade 3
	(1)	(2)	(3)	(4)
Treatment x End of Grade 3	0.081 [0.053]	0.029 [0.054]	0.063 [0.054]	-0.005 [0.013]
Treatment x Beginning of Grade 4	0.104* [0.055]	0.104 [0.066]	0.117* [0.063]	
Treatment x End of Grade 4	0.091** [0.044]	0.064 [0.049]	0.086* [0.047]	
Observations	6362	6362	6362	2391
p-value of equal coeffs.	0.930	0.672	0.739	
Control mean (non-std.)	11.49	8.778	0.393	0.129
Control S.D. (non-std.)	4.803	4.381	0.157	0.335

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 $H_0 : \theta_1 = \theta_2 = \theta_3$. Control mean and S.D. correspond to the baseline number of correct answers (cols. 1-3) or the proportion of students who repeated at the end of grade 3 (col. 4). 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 5: Dose-response Effects
Instrumental Variables Estimates, Standardized Outcomes

	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	0.009*** [0.002]	0.002 [0.001]	0.004*** [0.001]	0.000 [0.001]	0.007*** [0.001]
Days in tutorial x Beginning of Grade 4	0.007*** [0.001]	0.003** [0.001]	0.005*** [0.002]	0.004*** [0.001]	0.007*** [0.002]
Days in tutorial x End of Grade 4	0.007*** [0.001]	0.002** [0.001]	0.002 [0.001]	0.001 [0.001]	0.004*** [0.001]
Observations	4949	6362	6362	6362	6362
p-value of equal coeffs.	0.279	0.742	0.143	0.026	0.042
Baseline control mean (non-std.)	15.02	26.94	46.32	3.303	0.507
Baseline control S.D. (non-std.)	11.36	10.04	13.81	1.877	0.142

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 first stage Kleibergen-Paap F-statistic is equal to 1828 and the associated p-value is zero. The row labeled “p-value of equal coeffs.” shows the p-value of a test $H_0 : \beta_1 = \beta_2 = \beta_3$. Control mean and S.D. correspond to the number of correct answers (cols. 1-4) or the percent of correct answers (col. 5). 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 6: Targeting
(non-standardized outcomes)

	Non-words	Addition	Subtraction	Index
	(1)	(2)	(3)	(4)
Cohort 2	-4.024*** [0.718]	-2.151*** [0.296]	-1.425*** [0.267]	-2.533*** [0.348]
Cohort 3	-1.848*** [0.704]	-0.375 [0.291]	-0.057 [0.282]	-0.760** [0.355]
Observations	2610	2610	2610	2610
p-value of equal coeffs.	0.003	0.000	0.000	0.000
Cohort 1 control mean	28.33	12.23	9.212	16.59
Cohort 1 control S.D.	10.51	5.095	4.740	4.766

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%.

Table 7: Treatment Effect Heterogeneity
(standardized outcomes)

		Knowledge of letter sounds	Reading of non-words	Fluency of oral reading	Reading comprehen- sion	Literacy score
		(1)	(2)	(3)	(4)	(5)
Size of tutorial group	1–5 students	0.330*** [0.064]	0.002 [0.054]	0.063 [0.059]	−0.058 [0.065]	0.177*** [0.059]
	6 students	0.370*** [0.082]	0.152** [0.059]	0.254*** [0.062]	0.054 [0.044]	0.363*** [0.071]
	p-value of equal coeffs.	0.679	0.018	0.025	0.153	0.025
Peers' initial ability	High	0.366*** [0.066]	0.054 [0.063]	0.090 [0.064]	−0.065 [0.057]	0.230*** [0.068]
	Low	0.334*** [0.071]	0.096 [0.065]	0.221*** [0.064]	0.056 [0.052]	0.305*** [0.070]
	p-value of equal coeffs.	0.695	0.623	0.159	0.095	0.402
Homogeneity of tutorial group	s.d. below median	0.413*** [0.067]	0.097* [0.058]	0.202*** [0.060]	0.025 [0.052]	0.327*** [0.065]
	s.d. above median	0.283*** [0.068]	0.053 [0.058]	0.109* [0.057]	−0.034 [0.053]	0.207*** [0.063]
	p-value of equal coeffs.	0.092	0.511	0.237	0.385	0.107
Tutor's previous experience	No	0.468*** [0.101]	0.031 [0.064]	0.201*** [0.060]	0.041 [0.050]	0.313*** [0.076]
	Yes	0.245** [0.095]	0.252*** [0.080]	0.471*** [0.102]	0.098 [0.099]	0.458*** [0.095]
	p-value of equal coeffs.	0.066	0.017	0.018	0.578	0.161

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%.