Targeted Remedial Education: Experimental Evidence from Peru

Education Division

Juan E. Saavedra
Emma Näslund-Hadley
Mariana Alfonso

TECHNICAL NOTE Nº 1317

September 2017

Targeted Remedial Education: Experimental Evidence from Peru

Juan E. Saavedra*
Emma Näslund-Hadley**
Mariana Alfonso**

* University of Southern California, Dornsife Center for Economic and Social Research, and National Bureau of Economic Research.

**Inter-American Development Bank

September 2017

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

Saavedra, Juan E.
Targeted remedial education: experimental evidence from Peru / Juan E. Saavedra, Emma Näslund-Hadley, Mariana Alfonso.
p. cm. — (IDB Technical Note ; 1317)
Includes bibliographic references.
1. Remedial teaching- Education (Primary)-Peru. 2. Remedial teaching-Evaluation. 3. Public schools-Students-Scores. 4. Public schools-Examinations. I. Näslund-Hadley, Emma. II. Alfonso, Mariana. III. Inter-American Development Bank. Education Division. IV. Title. V. Series.
IDB-TN-1317

http://www.iadb.org

Targeted Remedial Education: Experimental Evidence from Peru

September 2017

Juan E. Saavedra† Emma Näslund-Hadley ‡ Mariana Alfonso ‡

Abstract*

Improving learning among low-achieving students is a challenge in education. We present results from the first randomized experiment of an inquiry-based remedial science education program for low-performing elementary students in a developing-country setting. Third-grade students in 48 low-income public elementary schools in Metropolitan Lima who score at the bottom half of their school distribution in a science test taken at the beginning of the school year are randomly assigned to receive up to 16 remedial science tutoring sessions of 90 minutes each. Control group compliance with assignment is close to perfect. Treatment group compliance is 40 percent, equivalent to 4.5 tutoring sessions, or a 4 percent increase in total science instruction time. Despite the low treatment intensity, students assigned to the remedial sessions score 0.12 standard deviations higher on a science endline test, with all gains concentrated among boys. We find no evidence of remedial education producing within-student spillovers to other subject areas (math or reading) or spillovers on other students in the classroom. We conclude that low-intensity remedial education can have an effect on science learning among low-achieving students.

---

* We thank Richard Murnane for helpful comments. We thank the Innovations for Poverty Action (IPA) Peru team, especially Andrea Cornejo and Adam Kemmis Betty for their invaluable support on the field. We acknowledge financial support from the Japan Poverty Fund of the Inter-American Development Bank. The research activities reported in this study have approval from the University of Southern California and IPA's Institutional Review Boards. This field trial was pre-registered with the American Economic Association's RCT Registry Number under code AEARCTR-0000379.

† University of Southern California, Dornsife Center for Economic and Social Research, and National Bureau of Economic Research, juansaav@usc.edu;

‡ Inter-American Development Bank

1. Introduction

Improving learning among low-achieving students is a challenge in education. Many developing countries—including most Latin American countries—have centralized education systems with standardized curricula and textbooks. Teachers typically teach the curriculum, rarely setting aside the time to assist struggling students (e.g. Banerjee et al 2007). Teachers may also lack skills to teach the same content at different skill levels to address the needs of all students (e.g. Duflo, Dupas and Kremer 2011).

Some argue that poor academic performance on international assessments may be the result of many students falling behind the national curriculum (e.g. Glewwe, Kremer, Moulin 2009). In Peru, for example, only 27 percent of grade 2 students meet the grade-specific proficiency level in mathematics (MINEDU 2015). In comparative international assessments, Peruvian students typically score at the bottom of the distribution. In the 2012 application of the PISA test, for instance, Peru ranked 65th in mathematics and 65th in natural science out of 65 participating nations (OECD 2013), and 75 percent of Peruvian students were low achievers in mathematics as compared to 23 percent for OECD students. In the 2013 TERCE regional study, close to 40 percent of 6^{th} grade Peruvian students scored at the lowest level of achievement in science (LLECE, 2015). To address these dismal results, Peru piloted in 2010 and 2012 two programs to improve science performance among third graders. These programs consisted mostly of training and coaching teachers to use student-center methodologies, and were based on the 2008 national curricular standards for science. Consistent with the idea of many students falling behind the national curriculum, a randomized evaluation concluded that the 2010 pilot program only improved science achievement of third-grade students with above-average baseline performance (Beuermann et al 2013).[1]

1 These results are consistent, for instance, with evidence from Kenya that textbook provision only raised the academic performance of high achievers at baseline (Glewwe, Kremer and Moulin 2009).

Remedial education, by which students receive targeted, self-paced teaching, shows promise at improving short- and medium-term academic performance of low-achieving students in a variety of contexts. The evidence on remedial education, however, is mostly limited to improving basic mathematics and literacy skills.[2] The evidence on remedial mathematics and literacy education suggests that direct instruction may be an effective pedagogical model for low-achieving students (e.g. Houtveen and van de Grift 2007, 2012; Kaiser et al 1989; Linan-Thompson and Vaughn 2007). However, research on whole-class science instruction suggests that inquiry-based instruction—in which students engage in hands-on practical work with different degrees of teacher guidance—improves learning more than traditional classroom practices (e.g. Brickman et al 2009; Ergül et al 2011; Hmelo-Silver 2004; Harris et al 2014).

It is unclear whether inquiry-based instruction is effective for low-achieving, early-grade students (e.g. Hmelo-Silver 2004). We present experimental evidence on an inquiry-based remedial science education program targeting low-achieving third-grade students in 48 low-income public elementary schools in Metropolitan Lima, Peru—those same students left behind by the previous pilot programs to improve science achievement among third graders. While the program is a typical tutoring intervention in that it provides additional instruction to small groups of students, to our knowledge, this

[2] In the US, for example, a meta-analysis of 31 elementary school tutoring programs for students at risk of reading failure concludes that tutoring raises reading achievement by .67 standard deviations (Elbaum, Vaughn, Hughes, and Moody 2000). Another US meta-analysis of 35 mathematics and reading tutoring interventions for at risk elementary students concludes that remedial education appears to be equally effective at improving learning in both subject areas (Lauer et al 2006). However, a meta-analysis of randomized experiments of elementary education volunteer tutors concluded that tutoring improves reading, but has no effect on math skills (Ritter et al 2009). In India, remedial tutoring targeting the bottom third of students raises student achievement, particularly in mathematics (Banerjee et al 2007). Evidence also suggests that trained teachers are more effective tutors than non-teachers. A systematic review of 97 tutoring interventions reveals that the impact on student achievement is greater with teacher tutors as opposed to volunteers and paraprofessionals (Slavin et al 2011). In terms of duration, a meta-analysis of 35 mathematics and reading tutoring programs concludes that programs of moderate duration (45–85 hours) have greater effect on both reading and mathematics achievement than programs of longer duration. The effects are similar whether tutoring took place in the afternoon, on Saturdays or during the summer (Lauer et al 2006).

is the first rigorous study to date to document impacts of an inquiry-based remedial science education program targeted at early grades. The inquiry-based approach is similar to the previous pilot programs, but differs in that students lagging behind are taught in smaller groups to their learning ability.

Students who score in the bottom half of their school distribution on a science test administered at the beginning of the 2014 school year are randomly assigned to receive up to 16 science remedial tutoring sessions of 90 minutes each or to not receive any type of remedial support (control group). The tutoring sessions follow an inquiry-based format and take place in schools—typically in the afternoon—in groups of nine students, on average. Tutors are public-sector elementary school teachers selected among volunteer candidates. Prior to the start of the tutoring sessions selected tutors received content knowledge and pedagogical training and received detailed and highly structured tutoring materials that included flipcharts with activities for each session and formative evaluation rubrics.

Control group compliance with assignment is close to perfect. Treatment group compliance, on the other hand, is only 40 percent, equivalent to 4.5 tutoring sessions. Despite the very low intensity of treatment, students assigned to remedial tutoring score 0.12 standard deviations higher on a science endline test. All the achievement gains, however, are concentrated among boys assigned to treatment, for which gains are 0.22 standard deviations.

These differential gains are not explained by gender differences in treatment compliance. A number of factors could have contributed to the concentration of gains entirely among boys. One possibility is that the absence of effect among girls stems from the preferential treatment of boys by tutors, with whom we observe they engage more proactively than with girls—even though the overwhelming majority of tutors in our study are women. Our observation of preferential treatment for boys is consistent with prior evidence documenting how stereotypical ideas held by teachers and differential teacher attention and effort devoted to boys versus girls perpetuate gender gaps in beliefs and competence in scientific endeavors (Mendick 2006; Fenema et al. 1990;). This finding may suggest that matching instructors' and students' genders does not necessarily help overcome STEM achievement gaps by gender (e.g. Carrell, Page and West 2010). Alternatively, the absence of effect among girls may stem from boys being more proactive than girls in small-group tutorials. This would be consistent with boys' monopolization of the science kits observed by Beuermann and colleagues (2013).

We find no evidence that inquiry-based science remedial tutoring assignment has within-student spillovers on achievement in other subject areas (math and literacy). We also find no evidence of spillovers on science achievement onto students from the same classrooms who do not receive the tutoring sessions.

The remainder of this paper is organized as follows. The second section discusses previous efforts in Peru to identify an effective primary education science model and the inquiry-based remedial science education we design and evaluate. The third section describes the sample and experimental design. The fourth section describes the data and analytical approach. The fifth section reports our main findings, and the sixth section concludes with a discussion of the limitations and implications of the results.

2. Background and Program Description

In this section, we describe science classroom practices and recent efforts to boost science skills in Peru that motivate the present study (subsection a) and describe the program we evaluate (subsection b).

a. Scientific learning classroom practices and efforts to boost science skills among Peruvian students

Peruvian students have poor overall performance in international assessments. In the last application of the PISA test, for example, Peru ranked 65th in mathematics, and 65th in natural science among the 65 participating nations (OECD 2013). One-third of Peruvian students place in the lowest proficiency level of PISA science, which means that they do not master even the most basic skills. The PISA assessment indicates that Peruvian students lack critical reasoning skills, the ability to analyze and synthesize information, and to apply new knowledge in real-life settings.

Lack of adequate teaching skills may help explain Peruvian students' poor performance in comparative science and math assessments. Teachers typically teach the curriculum without setting aside time for struggling students. While evidence suggests that Peruvian teachers overemphasize the least cognitively demanding topics and pose learning tasks that are not cognitively demanding, students rarely get teacher feedback and when they do, it is often erroneous (Cueto, Ramirez, and Leon 2006). Moreover, half of math teachers nationwide cannot perform basic arithmetical calculations (Alfonso, Bos, Duarte and Rondon 2012).

In Peru, scientific learning typically follows an explicit teacher instruction model (e.g. Kirschner 2006; Clark 2012). Teacher lectures take up most of class time and limited time is devoted to practical work. To the extent that they do, teachers conduct practical work themselves, limiting student opportunities for hands-on learning (Loera et al 2013; Näslund-Hadley et al 2014).

To address some of the country's math and science educational challenges, in 2010 Peru's government piloted a program that aimed to promote critical thinking and scientific reasoning skills among third-grade students. This program, based on the 2008 national curriculum that includes areas such as the physical world, the human body, and living beings and environment, aimed to teach children about scientific models and their applications. A key component of is the pilot science program was teacher training, with a specific focus on mastering the structure and content of inquiry-based learning approaches (e.g. Tutwiler and Grotzer 2013).[3]

A school-level randomized evaluation of the pilot teacher-training model in 62 districts of the state of Lima concluded that the program only improved science test-scores among boys in urban areas and students with above average baseline performance (Beuermann et al 2013). Adjustments were made to place an increased focus on girls' confidence in their science skills, and working groups were separated by gender for some activities to ensure that girls got hands-on experience. To close the geographical divide, in 2012 pilot efforts were made to increase compliance in rural areas to ensure that all teachers benefitted from mentoring. These adjustments to the program made the average gender and geographical gaps insignificant. However, among students in the bottom half of the baseline score distribution, the pilot program still had no impact (IPA 2014). As a consequence of the program, the science achievement gap between high and low performers widened.[4] These results motivate the present study, which investigates whether remedial tutoring sessions for the lowest performing students help reduce science achievement gaps between high- and low-performers.

3 Inquiry-based methods demand new teaching approaches and greater content knowledge on the part of the teacher, which may be challenging in some contexts (Clark et al 2012; Mayer 2004; Kirschner et al 2006). Among inquiry-based methods, a wide range of pedagogical approaches exists, ranging from structured inquiry to completely open inquiry (Colburn 2000). A meta-analysis of 37 experimental and non-experimental studies of inquiry-based instruction concludes that learning is optimized when teachers guide inquiry rather than students engaging in completely open inquiry (Furtak et al 2012).

4 These results are consistent, for instance, with evidence from Kenya that textbook provision only raised the academic performance of high achievers at baseline (Glewwe, Kremer and Moulin 2009).

b. Program Description: The Science Remedial Tutoring Program

The Science Remedial Tutoring Program aims to help low-performing students master theoretical and practical knowledge related to science through inquiry-based methods. The goal is that, when confronted with an unfamiliar situation, students are able to develop relevant answers through critical thinking and collaborative work. As a by-product, the program seeks to promote healthy study habits, academic motivation and love of learning.

Universidad Cayetano Heredia developed the structure and contents of the Science Remedial Tutoring Program. The program has four components: (i) development of pedagogical materials; (ii) selection and training of tutors; (iii) selection of students, and iv) implementation of tutoring sessions in schools.

c. Pedagogical materials

To develop the pedagogical materials, Universidad Cayeteno Heredia employed two local pedagogy specialists, one specialist in primary education and one in science education. These two specialists developed the tutoring materials, the contents of which are based on the 2008 National Curricular standards for teaching science to third-graders.

Based on the National Curriculum, to bridge the issue of gaps in tutors' content knowledge, the specialists developed detailed and highly structured tutoring materials that included flipcharts with activities for each session and formative evaluation rubrics. That is, the materials combine elements of explicit instruction with inquiry-based activities. In this inquiry-based approach, tutoring sessions begin with a challenge/question. For example, as part of a weather module, students explored why Lima is covered in fog. The tutor guided them in the formulation of hypotheses, design of experiments, and discussion of their findings as the students made their own fog in jars.

Students were then encouraged to formulate preliminary answers based on prior knowledge, acquire new information through experimentation and reading, re-structure prior knowledge, establish conclusions and apply the new knowledge to unfamiliar situations.

d. Selection and training of tutors

For the Science Remedial Tutoring Program, Universidad Cayetano Heredia selected 16 tutors –15 of which were women, like the majority of public school teachers in Peru. Universidad Cayetano Heredia chose a male to tutor students in schools located in high-crime areas. Tutor selection took place between March and May 2014. Selection criteria included: (i) minimum two years of primary school teaching experience; (ii) positive attitude towards the teaching and learning of science; (iii) assertive communication and class-management skills; and (iv) ability to create respectful, empathetic and tolerant relationships with children.

Tutors are local primary or secondary public school teachers, although not necessarily teachers in the schools in which they provide tutoring. Tutors are paid an hourly wage of US$10 for their services and transportation, which is slightly below than what primary education teachers earn on average (US$14 per hour). The pay was not linked to the performance of the tutor, nor were there any prospect of continued employment after the program ended. Tutors are assigned to target tutoring schools based on geographic proximity to their residential location.

Once selected, tutors participated in a training workshop organized by Universidad Cayetano Heredia. The two education specialists led the workshop. The workshop took place before the start of the 2014 school year and lasted 20 hours, split over six days. The general goal of the workshop was to train tutors in the pedagogical and didactical foundations of inquiry based learning. As such, tutors were encouraged to apply in each session seven principles: (i) learning builds on prior knowledge; (ii) learning is a restructuring of prior knowledge; (iii) learning takes place in the interaction with the object of study; (iv) learning requires language and communication; (v) emotions affect learning; (vi) learning is a social process as well as a psychological process; and (vii) learning requires self-regulation (meta-cognition).

Tutors were instructed on possible approaches to apply these foundational principles to each of the tutoring activities in order to engage students. Some of these approaches include encouraging and discussing different points of view, sequencing contents to follow the children's logic and applying new perspectives to unfamiliar situations. In the workshop, the specialists and tutors also reviewed the content and activities for each session.

During the workshop, tutors received an instructional guide summarizing principles, pedagogical approaches and activities for each tutoring session. The two specialists also provided ongoing support to tutors during the implementation of the tutoring sessions.

e. Selection of students

The program targeted low-performing science third-grade students in 48 public elementary schools in Metropolitan Lima (sample selection details below). Baseline performance was assessed through a written test administered during class in May 2014. Within each school, the program targeted the bottom 50 percent of scorers. Eligibility for tutoring participation was then determined using a lottery assignment mechanism (details below).

f. Implementation of tutoring sessions

Tutoring sessions took place in each of the 48-participating public elementary schools in Metropolitan Lima. There was a total of 70 tutoring groups. Each tutor was assigned on average to 5 tutoring groups (some as few as 3 and some as many as 7). There were more tutoring groups than schools because some of the schools had very large third-grade classes or more than one third-grade section. Anywhere between 3 and 17 students were assigned to each tutoring group (always in the school they attended), with mean group size of 9 students.

Tutoring began in July 2014, halfway through the Peruvian school year, which begins in March. Tutoring consisted of 16 weekly 90-minute sessions and ended in November 2014. In total, selected students could receive up to 24 hours of additional tutoring, a 14 percent increase in total instructional time relative to the regular science schedule.

Tutoring sessions took place at each school's premises. Most tutoring sessions were scheduled in the afternoon (at the end of the school day). In a few cases, sessions were scheduled in the morning (for students attending school in the afternoon) or on Saturday. In the first session students received a workbook called "Making and Learning Science," which describes various scientific inquiry activities that students could pursue independently.

Each tutor was responsible for coordinating and scheduling sessions with her groups. Tutors initially approached school principals and third-grade teachers to explain program details seeking support to promote attendance of eligible students. Tutors also invited parents of eligible students to tutoring information sessions during which the goals of the tutoring program, the approach, and the expected benefits were explained.

Parent attendance to the tutoring information sessions was low and decreased particularly towards the end of the school year. To ensure that all parents were informed about the availability of the tutoring program, students were also asked to bring home an information sheet that parents were supposed to sign and send back. Some tutors also visited the students' homes aiming to contact parents. In total, about 50 percent of the parents of students assigned to tutoring signed and returned these forms. This suggests that at least 50 percent of parents knew about the availability of the program for their children. The take up rate at the student level is discussed below.

3. Evaluation Sample, Experimental Design and Randomization Balance

a. Evaluation Sample

We collected baseline test score data to determine eligibility for the Science Remedial Tutoring Program from third-graders in 51 public elementary schools in Metropolitan Lima in May 2014. Of these 51 schools, 39 had participated in the 2012 Science Education Teacher-Training program. We chose these 39 schools to facilitate access to the tutors, as these schools had prior contact with the training staff from Universidad Cayetano Heredia. The remaining 12 schools were randomly chosen among comparable schools in the poorest localities in Metropolitan Lima. After baseline data collection, we discarded two schools because they had less than 8 third-grade students, and we wanted to minimize the risk of stigmatizing one or two students with eligibility for participation. We further discarded one school because we were unable to contact tutoring-eligible children. The final evaluation sample is, therefore, drawn from the remaining 48 public elementary schools in Metropolitan Lima.

The typical school in the evaluation sample has two third-grade sections and 51 third-grade students. The principal of the average school in the sample has 6.3 years of experience as school principal and teachers have 5.6 years of experience, 4.6 of which are in the current school. About 14 percent of sample teachers had participated in the 2012 Science Education Teacher-Training Program. About 55 percent of students in the sample are boys and the average student age is just over 8 years old.

b. Experimental Design

We collected baseline science test-score data in May 2014 from 2,399 third-grade students in the 48 schools of the evaluation sample. The science, mathematics and Spanish language tests were simplified versions of the tests administered as part of the evaluations of the 2010 and 2012 science pilot programs implemented in Lima. The tests were developed to measure third grade skills based on Peru's 2008 new basic education curriculum and national study plan for third grade mathematics, Spanish and science. In science, the third-grade curriculum includes the Physical World and Preservation of the Environment; the Human Body and Health; and Animals and their Environment.

Test questions addressed a mixture of content and critical thinking skills. Content questions included, for example, questions about how different food groups can help us stay healthy; and the identification of Peruvian animals. As an example of a critical thinking question, students were asked why a snow cone turned into red water when a little girl left it on a bench while playing. A supervisor monitored and timed the students as they individually completed the learning test in writing.

The bottom 50 percent of scorers in the baseline test within each school was targeted for tutoring. In total, 1,219 students were targeted for tutoring. Among these 1,219 students, we randomly assigned eligibility to participate in the Science Remedial Tutoring Program stratifying by school and gender. In practice, we only had 95 lotteries (48x2 – 1) because in one school only boys scored in the bottom 50 percent of the school test-score distribution. In the final evaluation sample, we have 609 students assigned to treatment (331 boys and 278 girls) and 610 students assigned to control conditions (337 boys and 273 girls)—that is, they did not receive remedial tutoring.

c. Randomization Balance

Based on the baseline test and a socio-demographic questionnaire, randomization balanced characteristics across students assigned to treatment and to control (Table 1). These characteristics include gender composition, age, school shift (morning or afternoon), whether the student is Spanish speaking, the number of adults in the household, whether the father of the student is present in the household and baseline test-scores (Panel A, Table 1). At baseline, boys and girls score at comparable levels in science, math and reading. The only statistically significant (at the 10 percent) baseline difference in means in favor of the treatment group is math scores among boys (Panel B, Table 1).

4. Data and Empirical Strategy

a. Data

We use three data sources to document impacts of the Science Remedial Tutoring Program. The first data source is the baseline test and socio-demographic questionnaire collected from third-grade students in the 48 schools in the sample.[5] The second data source is data on student attendance to the tutoring sessions collected by the tutors (i.e. compliance with treatment assignment). These data were collected for 12 out of 16 tutoring sessions. For the first 4 sessions tutors did not collect student attendance data. We measure student attendance three different ways: (i) number of tutoring sessions attended and (ii) fraction of tutoring sessions attended and iii) total additional minutes spent in tutoring. The final data source is endline test and student survey data, collected in November 2014, about 5 months after the start of the tutoring sessions. Based on evidence from mathematics and reading tutoring programs, programs of moderate

5 The questionnaire was designed and validated for use among elementary students. For example, it did not include questions about income, but rather about the dwelling of the household.

duration (45–85 hours) have greater effect on both reading and mathematics achievement than programs of longer duration (Lauer et al 2006).

Endline attrition was low, uncorrelated with treatment status and with the demographic composition of randomization groups (Table 2). Over 90 percent of students assigned to control took the endline test (Bottom of Column 1, Table 2). Students assigned to treatment are 1 percentage point more likely to take the test but the difference is not statistically significant, with or without baseline controls (Columns-1-3, Table 2). Column 4 of Table 2 shows that the demographic composition of randomization groups at endline is also balanced, since the F-statistic of the joint test of interactions between baseline characteristics and treatment assignment is 1.50 (p-value= 0.19).

b. Empirical Strategy

In all tables, we begin by showing unadjusted mean differences in outcomes between students assigned to science remedial tutoring and to the control group. Our preferred models, however, are test-score value-added specifications of the following form:

$$Y_{isj,t} = \alpha + \tau D_i + \beta Y_{isj,t-1} + \gamma X_i + \theta_s + \varepsilon_{isj,t}$$

where $Y_{ijs,t}$ is the endline score for student *i* in school *s* and subject *j*; D_i is tutoring assignment status, which equals one if student *i* is randomly assigned to treatment and zero if not; $Y_{isj,t-1}$ is the baseline score for student *i* in school *s* and subject *j*; X_i are student socio-demographic characteristics measured at baseline described in Table 1; θ_s are school fixed effects to account for the stratified randomization design, and $\varepsilon_{isj,t}$ are error terms that we allow to arbitrarily co-vary within schools. The key coefficient of interest is τ, which captures the impact of being assigned to receive remedial science tutoring on test-scores, in other words, an Intent-to-treat effect. In all tables we add the control variables sequentially to document the stability of estimates of τ to alternative sets of control variables. We also explore heterogeneity by gender in the estimates of τ.

5. Results

In this section, we discuss results on compliance with treatment assignment (subsection a), endline science achievement (subsection b), endline Math and Reading achievement (subsection c), heterogeneity by gender (subsection d) and spillovers to students in the treatment schools who were not treated (subsection e).

a. Compliance with treatment assignment: Attendance to Remedial Science Education sessions

Compliance with treatment assignment among students assigned to control was very high, so that control group contamination was negligible. On average, students assigned to control conditions attended 0.04 tutoring sessions (Panel A, Table 3) or alternatively, received three additional minutes of total tutoring time (Panel C, Table 3).

Compliance with treatment assignment among students assigned to the remedial treatment, however, was low. For the 12 tutoring sessions for which we collected attendance data, students assigned to treatment attended, on average, 4.7 sessions (Panel A, Table 3).[6] This estimate on attendance represents a compliance rate among students eligible for treatment of 39 percent (Panel B, Table 3). Students assigned to tutoring received, on average, 430 minutes of additional remedial science tutoring (Panel C, Table 3). This implies that rather than the intended additional 1,440 minutes (14 percent) of science instruction time, the average student received an additional 4 percent of science instruction time. Baseline science scores and students' gender are

[6] The number of sessions attended is likely a lower bound if students assigned to treatment were also more likely to attend the initial four sessions for which we do not have attendance data.

uncorrelated with tutoring attendance (Columns 2-4, Table 3). The low attendance is the result of a combination of factors, including failure to effectively promote the program and its benefits among students and parents. Students may also have time conflicts with other responsibilities as 43 percent of Peruvian 5 to 17-year-olds are economically active, generally combining school with work. Although child labor is 40 percent more prevalent in rural areas, also children in urban areas are economically active, mainly as street vendors (ILO, 2009). The prevalence of child labor increases with age, which would be consistent with our finding that younger students eligible for treatment were more likely to attend remedial tutoring (Column 4, Table 3). Moreover, children may need to help at home in the afternoon or during weekends by taking care of younger siblings while their parents are working. Attendance diminished toward the end of the year as a result of competing extracurricular activities.

b. Endline science test scores

Remedial tutoring assignment increases endline science scores. When measured in percentiles of the test score distribution, the estimate of the impact of treatment assignment is between 3 and 4 percentiles. The ITT estimates are robust to the inclusion of alternative sets of control variables (Columns 1-4, Table 4). When measured in standard deviation units, the estimate of treatment assignment is between 0.12 and 0.14 standard deviations (Columns 5-8, Table 4).

ITT impact estimates of remedial tutoring are substantial in magnitude once we account for the intensity of treatment among students assigned to tutoring. For example, taking at face value the 39 percent difference in compliance rate between students assigned to tutoring and control (Panel B, Table 3) would imply estimates for the treatment on the treated of 0.30-0.36 standard deviations. We cannot, however, give a causal interpretation to this Wald estimate because doing so would require assuming that tutoring assignment only affects student outcomes through participation in the remedial tutoring sessions. This assumption may be challenged if, for example, non-participants benefit indirectly through improved regular classroom learning as a result of a lower fraction of underperforming students delaying the pace of learning. We explore the empirical support for these potential spillover effects in subsection (e) below.

The gains of tutoring assignment on endline science achievement accrue to students who scored at baseline between negative one standard deviation and average score (Figure 1). In the sample, this corresponds roughly to students at or above the 10th percentile. This result suggests that while the benefits of science tutoring are fairly widespread through the baseline achievement distribution of low-performing students, the Science Remedial Tutoring program is ineffective at improving the achievement of the lowest 10 percent of students.

c. Endline test scores on other subjects: Math and Reading

Science remedial tutoring assignment appears to shift the distribution of Math and particularly Reading scores for students in the middle of the distribution (Figure 2). However, point estimates of the effect of remedial tutoring assignment on endline Math achievement tests are often negative, always small in magnitude compared to those on the science test and never statistically significant (Panel A, Table 5). For Reading, impact estimates are positive and although small, we nevertheless cannot reject the hypothesis that they are comparable in magnitude to those on the science test (Panel B, Table 5). While this evidence is consistent with the possibility that the inquiry-based approach used in the Science Remedial Tutoring program has impact in other subject areas, the effect is small.

d. Heterogeneity by gender

The effects of the Science Remedial tutoring program on endline science achievement are entirely driven by gains among males. Tutoring assignment increases science scores for boys by about 5 percentiles (Columns 1-4, Table 6) or about 0.22 standard deviations (Columns 5-8, Table 6). For girls, tutoring impacts are negligible. Coefficients on the interaction term are about -4.5 percentiles or -0.21 standard deviations. Estimates of coefficient on the interaction term are generally statistically significant at the 10 percent level (5 percent with full controls with the dependent variable expressed in standard deviations, Column 8, Table 6).

One possible explanation to the impact heterogeneity by gender is differences in treatment intensity (compliance) between boys and girls. We do not find empirical support for this conjecture. Boys and girls are equally likely to attend tutoring sessions; the coefficients on the interaction term are always small relative to the main treatment effect and never statistically significant (Table 7). The literature on gender and STEM education may point to a different explanation for the concentration of gains entirely among boys. Stereotypical ideas held by teachers and differential teacher attention and effort devoted to boys versus girls have been shown to perpetuate gender gaps in beliefs and competence in scientific endeavors (Mendick 2006; Fenema et al. 1990). Based on this literature, it is possible that the tutors displayed a preferential treatment towards boys and more proactively engaged with boys than with girls during the inquiry-based activities. We cannot, however, rule out that boys were simply more actively engaged during the tutoring sessions and monopolized the science kits, which would also be consistent with this literature.

e. Spillovers

Since we randomized at the student level within schools it is possible that tutoring assignment generated spillovers on other students within the classroom that were not eligible to participate. For example, non-participants may have benefited indirectly through improved regular classroom learning as a result of a lower fraction of underperforming students delaying the pace of learning.

While we cannot use purely experimental variation to test for the presence of spillovers on other students, our research design creates variation within classrooms in the fraction of students receiving treatment. This is so because our randomization stratifies treatment assignment by school and gender but not by classroom. For schools with more than one third-grade section, this design therefore generates variation in the fraction of students assigned to remedial tutoring within a classroom.

Figure 3 shows that this variation is considerable. In some sections, no students are assigned to receive remedial tutoring. There are a number of sections in which anywhere between 20 and 60 percent of students are assigned to tutoring. In one section, all students are assigned.

In this subsection, we take advantage of this variation to estimate learning spillovers of tutoring. Specifically, under a linear-in-means peer-effects model, if (positive) spillovers exist, student achievement should be higher in sections with a higher fraction of students assigned to remedial tutoring. To test this hypothesis, we use the following regression model:

$$Y_{isc,t} = \alpha + \pi\overline{D}_c + +\delta\overline{Y}_{sc,t-1} + \beta Y_{isc,t-1} + \gamma X_i + \theta_s + \varepsilon_{isc,t}$$

where $\overline{D_c}$ is the section-level fraction of students assigned to remedial tutoring and $\overline{Y}_{sc,t-1}$ is the same-subject average baseline performance of section *c* in school *s*. In this specification, we include baseline section-average performance; without it, $\overline{D_c}$ is mechanically (negatively) correlated with the outcome variable. This is so because tutoring targets low-performing students, so a high fraction of students assigned to tutoring in a section implies a high fraction of low-performing students in the section. By including the section-average baseline performance and school fixed effects, the thought experiment we have in mind asks whether among two sections of the same school with similar baseline composition, students in the section with a higher fraction of students assigned to tutoring perform better than students in the section with a lower fraction of tutoring-assigned students. As before, we also allow error terms to arbitrarily co-vary within schools.

We find no evidence of tutoring learning spillovers among other students in the same section. Without controls the fraction of students assigned to tutoring in a section is negatively correlated with endline science achievement (Column 1, Table 8). Including the section average and baseline individual achievement flips the sign (Column 2, Table 8). However, estimates in Column 2 also rely on across school variation in section composition, which is problematic. When we include school fixed effects and full demographic controls the coefficient on the fraction of student assigned to tutoring in a section is small, negative and not statistically significant (Columns 3 and 4, Table 8).

6. Conclusion

Society and economies benefit from a scientifically literate population. Considering Latin America's meager results on international standardized science assessments, it is important to identify learning models that help ensure that all children boost their scientific skills. The 2010 and 2012 primary Science education pilots implemented in Peru helped improve science learning. Yet, the pilots revealed that one size does not fit all. Although the learning models that were piloted provided differentiated instruction according to the abilities of diverse groups of students, they were not effective for students who scored at the bottom of the distribution in the baseline assessment. These students needed extra help in the areas where they were struggling.

There is a growing literature on the effectiveness of remedial education, showing that remedial education based on direct instruction tends to be more effective than unstructured pedagogical approaches (e.g. Houtveen and van de Grift 2007, 2012; Kaiser et al 1989; Linan-Thompson and Vaughn 2007). However, there is scant literature on remedial natural science education in the elementary grades. In the absence of rigorous evaluations of remedial science education in the elementary grades, we turned to research on what works in regular classroom science instruction in the elementary grades, which points to the effectiveness of inquiry-based classroom practices. This raises the question if inquiry-based approaches can effectively be used also in remedial education to improve learning among low-performing students.

The results we present here are the first that measure a science education program targeted at lower performing students in early grades. It is also the first randomized experiment of a science tutoring program for small groups of lower performing students in Latin America.

Our results indicate that, despite a fairly low treatment intensity due to low (one-sided) compliance among students assigned to treatment, assignment to treatment increased science achievement by between 0.12 and 0.14 standard deviations. These results are striking because they were achieved among a group of students who did not improve as a result of the initial universal interventions for all students.

These findings suggest that low-performing students can learn through inquiry-based pedagogical approaches. This remedial science education model could easily be expanded to provide intensive academic support at a large scale for students who fall behind. Since the tutors are local and the training is short, the project would be straightforward to replicate. However, to bring this remedial science education model to scale, we identify two important challenges. First, the remedial inquiry-based Science education program did not significantly improve learning among girls. A challenge remains to identify instructional models that help mitigate instructor stereotypes and attitudes with regards to gender and science. Second, the overall effectiveness of the remedial education model was achieved in spite of a very low compliance rate. The effect could potentially be increased if the compliance rate increased. The compliance rate may improve by more clearly disseminating the program and promoting its benefits among parents and students. Since many students are either economically active or provide help at home by taking care of younger siblings, a more flexible tutoring schedule could also help improve the compliance rate.

In terms of directions for future research, two areas stand out. First, although our findings indicate that an inquiry-based tutoring model can improve learning among low-performing students in the early grades, additional research would be required to determine if this approach is more effective than tutoring based on a direct instruction methodology.

Second, despite the strong average effect, the inquiry-based tutoring model did not improve learning among the very bottom performers. This is consistent with previous literature that point to the challenge of improving the performance of children with severe learning disabilities (Torgesen 2001). Future research is needed to explore what models of remedial instruction can help accelerate learning in this group of students.

References

Alfonso, Mariana, Maria Soledad Bos, Jesus Duarte and Carlos Rondon. 2012. Panorama General de la Educación en América Latina y el Caribe, in *Educación para la Transformación*, Marcelo Cabrol y Miguel Szekely (Eds.). IDB. Washington, DC.

Banerjee Abhijit, Cole Shawn, Duflo Esther, Linden L. 2007. Remedying education: evidence from two randomized experiments in India. *Quarterly Journal of Economics,* 122:1235–64

Beuermann, Diether, Emma Näslund-Hadley, Inder J. Ruprah and Jennelle Thompson. 2013. The Pedagogy of Science and Environment: Experimental evidence from Peru. *The Journal of Development Studies*. 49(5): 719-736.

Brickman Peggy, Cara Gormally, Norris Armstrong and Brittan Hallar. 2009. Effects of inquiry-based learning on students' science literacy skills and confidence. *International Journal of Scholarship of Teaching and Learning*, 3(2): Article 16.

Carrell, Scott E., Marianne E. Page and James E. West. 2010. Sex and Science: How professor Gender Perpetuates the Gender Gap. *The Quarterly Journal of Economics*, 125(3): 1101-1144.

Clark, Richard E., Paul A. Kirschner and John Sweller. 2012. Putting Students on the Path to Learning: The Case for Fully Guided Instruction. *American Educator*, Spring 2012, Washington, DC: AFT.

Colburn, Alan. 2000. An Inquiry Primer. *Science Scope,* 23(6): 42-44.

Cueto, Santiago, Cecilia Ramirez, and Juan Leon. 2006. Opportunities to Learn and Achievement in Mathematics in a Sample of Sixth Grade Students in Lima, Peru. *Educational Studies in Mathematics* 62 (1): 25–55.

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

Elbaum, Batya, Sharon Vaughn, Marie Tejero Hughes, and Sally Watson Moody. 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): 605-619.

Ergül, Remziye, Yeter Simsekli, Sevgül Calis, Zehra Özdilek, Sirin Göcmencelebi and Meral Sanli. 2011. The Effects of Inquiry-based Science Teaching on Elementary School Students' Science Process Skills and Science Attitudes. *Bulgarian Journal of Science and Education Policy*, Volume 5(1): 48-68.

Fenema, Elizabeth, Penelope L. Peterson, Thomas Carpenter, and Cheryl Lubinski. 1990. Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics*, 21: 55-69.

Furtak, Erin M., Tina Seidel, Heidi Iverson, and Derek C. Briggs. 2012. Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. Review of Educational Research, 82 (3), 300–329.

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

Harris, Christopher J., William R. Penuel, Angela Haydel DeBarger, Cynthia D'Angelo and Lawrence P. Gallagher. (2014). *Curriculum materials make a difference for next generation science learning: Results from year 1 of a randomized controlled trial.* Menlo Park, CA: SRI International.

Hmelo-Silver, Cindy E. 2004. Problem-Based Learning: What and How do Students Learn? *Educational Psychology Review, 16* (3).

Houtveen, Thoni and Wim van de W. Grift. 2007. Reading Instruction for Struggling Learners. *Journal of Education for Students Placed at Risk*. 12(4), 405-424.

Houtveen, Thoni and Wim van de W. Grift. 2012. Improving Reading Achievements of Struggling Learners. *School Effectiveness and School Improvement*, 23(1): 71-93.

IPA (Innovations for Poverty Action). 2014. Educación de Ciencias II: Evaluación experimental del Programa Piloto de mejora de la Educación de Ciencias de la Universidad Cayetano Heredia en el Departamento de Lima, Perú. IDB Unpublished Consultancy Report, Lima: Peru.

ILO (International Labor Organization). 2009. *Trabajo Infantil en el Perú: Magnitud y perfiles vulnerables. Informe Nacional 2007-2008*. Peru.

Kaiser, S., Palumbo, K., Bialozor, R. C., & McLaughlin, T. F. 1989. The Effects of Direct Instruction with Rural Remedial Students: A brief report. *Reading Improvement* (26) 88-93.

Kirschner, Paul. A., John Sweller, and Richard E. Clark. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41, 75–86.

Lauer, Patricia A., Motoko Akiba, Stephanie B. Wilkerson, Helen S. Apthorp, David Snow and Mya L. Martin-Glenn. 2006. Out-of-school time programs: A meta-analysis of effects of at-risk students. *Review of Educational Research* (76): 275–313.

Linan-Thompson, Sylvia and Sharon Vaughn. 2007. Research-based Methods of Reading Instruction for English Language Learners. Alexandria, VA: Association for Supervision and Curriculum Development.

LLECE (Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación). 2015. Informe de Resultados Tercer Estudio Regional Comparativo: Factores Asociados. Santiago: UNESCO.

Loera, Armando, Emma Näslund-Hadley and Haydee Alonzo. (2013). El desempeño pedagógico de docentes en Nuevo Leon: Hallazgos de un estudio basado en videos de lecciones de matemáticas y ciencias. *Revista Latinoamericano de Estudios Educativos*. Vol. XLIII, No. 2, 11–41.

Mayer, Richard E. 2004. Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*. 59, 14–19.

Mendick, Heather. 2006. Masculinities in Mathematics. Maidenhead: Open University Press McGraw-Hill Education.

MINEDU 2015. Informe Nacional, Evaluación Censal de Estudiantes. Lima, Peru: Ministry of Education.

Näslund-Hadley, Emma, Armando Loera Varela and Ketelyn Hepworth. 2014. What goes on inside Latin American Math and Science Classrooms: A video study of teaching practices. *Global Education Review*. 1(3): 110-128.

OECD (Organization of Economic Cooperation and Development). 2013, OECD Fact Book 2013: Economic, Environmental and Social Statistics, OECD Publishing, Paris.

Ritter, Gary W., Joshua H. Barnett, George S. Denny and Ginger R. Albin. 2009. The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research*, 79(1): 3-38.

Slavin, Robert E., Cynthia Lake, Susan Davis and Nancy A. Madden. 2011. Effective programs for struggling readers: A best evidence synthesis. *Educational Research Review*, 6(1): 1-26.

Torgesen, Joseph K., Ann W Alexander, Richard K. Wagner, Carol A. Rashottte, Kytja K.S. Voeller, and Tim Conway. 2001. Intensive Remedial Instruction for Children with Sever Reading Disabilities: Immediate and Long-term Outcomes from Two Instructional Approaches. *Journal of Learning Disabilities*. 34(1): 33-58.

Tutwiler, M. Shane and Tina Grotzer. 2013. Why immersive, interactive simulation belongs in the pedagogical toolkit of Next Generation science: Facilitating student understanding of complex causal dynamics. In I. Saleh (Ed.) *Italic/Approaches and Strategies in Next Generation Science Learning,* Hershey, PA: IGI Global.

Tables and Figures

Table 1. Randomization Balance

	Assigned to Remedial Tutoring	Assigned to Control	Difference T/C	t-statistic (absolute value)
A. Socio-demographic Characteristics				
Female	0.46	0.45	0.01	0.31
Age	8.20	8.19	0.01	0.32
Morning shift	0.90	0.91	-0.01	0.68
Spanish-speaking	0.88	0.89	-0.01	0.54
Number of Adults in Household	2.39	2.37	0.03	0.36
Father Present	0.83	0.83	0.00	0.22
B. Baseline Scores (Raw percent of correct answers)				
B.1. Full Sample				
Science	0.51	0.51	0.00	0.02
Math	0.63	0.61	0.02	1.18
Reading	0.77	0.78	0.00	0.22
B.2. Boys				
Science	0.50	0.50	0.00	0.39
Math	0.63	0.60	0.03	1.67*
Reading	0.76	0.75	0.01	0.51
B.3 Girls				
Science	0.52	0.52	0.00	0.38
Math	0.63	0.63	0.00	0.04
Reading	0.79	0.81	-0.02	1.01
Observations	609	610		

Notes: Table shows results of raw mean comparisons (i.e. not adjusting for the stratified research design) across students assigned to Remedial Tutoring and to control conditions. Sample is 1,219 third-grade students who score in the bottom 50 percent of the baseline science test administered in May 2014 to third-grade students in 48 public elementary schools in Metropolitan Lima.

Table 2. Endline Attrition Analysis

	Dependent Variable is 1 if Student Took Endline Test and 0 if not			
	(1)	(2)	(3)	(4)
Assigned to Remedial Tutoring	0.011	0.011	0.012	0.329
	(0.016)	(0.016)	(0.015)	(0.180)
Baseline Science Score (s.d.)		0.016	0.014	0.02
		(0.007)*	(0.008)	(0.013)
Female			-0.017	-0.032
			(0.014)	(0.025)
Age			-0.053	-0.04
			(0.010)**	(0.014)**
Morning shift			-0.041	-0.046
			(0.031)	(0.030)
Spanish Speaking			-0.021	-0.021
			(0.019)	(0.028)
Number of Adults in Household			-0.001	0.002
			(0.006)	(0.008)
Father present			0.022	0.07
			(0.025)	(0.039)
Female*Assigned				0.034
				(0.035)
Age*Assigned				-0.029
				(0.021)
Morning Shift*Assigned				0.010
				(0.074)
Spanish Speaking*Assigned				0.000
				(0.042)
Number of Adults*Assigned				-0.007
				(0.015)
Father Present*Assigned				-0.105
				(0.049)*
Baseline Science Score*Assigned				-0.012
				(0.015)
Control Group Mean	0.921			
F-stat of joint hypothesis that interaction terms are all zero				1.50
p-value of F-stat				0.19
N	1,219	1,219	1,219	1,219

Notes: Table shows attrition analysis across students assigned to Science Remedial Tutoring and to control conditions. Sample is third-grade students who score in the bottom 50 percent of the baseline science test administered in May 2014 to third-grade students in 48 public elementary schools in Metropolitan Lima. * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Table 3. Compliance with Treatment Assignment: Student Attendance to Remedial Tutoring

	(1)	(2)	(3)	(4)
A. Number of Tutoring Sessions Attended				
Assigned to Remedial Tutoring	4.741	4.74	4.751	4.76
	(0.447)***	(0.447)***	(0.462)***	(0.458)***
Baseline Science Score (Percentile)		0.001	-0.003	-0.004
		(0.009)	(0.004)	(0.004)
Female				-0.023
				(0.119)
Age				-0.354
				(0.102)***
Control Group Mean	0.036			
B. Percent of Tutoring Sessions Attended				
Assigned to Remedial Tutoring	38.848	38.842	38.932	38.993
	(3.590)***	(3.592)***	(3.717)***	(3.684)***
Baseline Science Score (Percentile)		0.018	-0.023	-0.03
		(0.076)	(0.033)	(0.033)
Female				0.066
				(0.961)
Age				-2.625
				(0.797)***
Control Group Mean	0.301			
C. Additional Tutoring Time (Minutes)				
Assigned to Remedial Tutoring	430.406	430.346	432.87	433.726
	(49.457)***	(49.494)***	(51.186)***	(50.811)***
Baseline Science Score (Percentile)		0.161	-0.501	-0.604
		(0.965)	(0.401)	(0.410)
Female				2.167
				(10.888)
Age				-29.651
				(9.765)***
Control Group Mean	3.241			
N	1,219	1,219	1,219	1,219

Notes: Standard Errors clustered at the school level in parentheses. (1) No controls; (2) Controls for baseline science scores (percentile); (3) Controls for baseline science scores and school fixed effects; (4) Controls for baseline scores, school fixed effects and other student socio-demographic characteristics not shown in the table including school shift, Spanish speaking, adults in household and father present in household. * *p*<0.1; ** *p*<0.05; *** *p*<0.01.

Table 4. Remedial Tutoring Impacts on Endline Science Test Scores

	Dependent Variable is Test Score Percentiles				Dependent Variable is Test Score Standard Deviations			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Assigned to Remedial Tutoring	3.708	3.612	3.265	3.357	0.132	0.139	0.121	0.124
	(1.190)***	(1.203)***	(1.236)**	(1.256)**	(0.050)**	(0.051)***	(0.052)**	(0.053)**
Baseline Science Score		0.641	0.555	0.550		0.449	0.389	0.383
		(0.071)***	(0.052)***	(0.054)***		(0.049)***	(0.036)***	(0.035)***
Female				-0.316				0.009
				(1.371)				(0.055)
Age				-1.922				-0.09
				(0.851)**				(0.032)***
Control Group Mean	34.66				0.000			
	(1.873)***				(0.076)			
N	1129	1129	1129	1129	1129	1129	1129	1129

Notes: Standard Errors clustered at the school level in parentheses. Table shows science endline impact results. In columns (1)-(4) outcome variable and lagged test-score regressor are expressed in percentiles. In columns (5)-(8) outcome variable and lagged test-score regressor are expressed in standard deviation units. (1) and (4) No controls; (2) and (5) Controls for baseline science score; (3) and (6) Controls for baseline science scores and school fixed effects; (4) and (8) Controls for baseline scores, school fixed effects and other student socio-demographic characteristics not shown in the table including school shift, Spanish speaking, adults in household and father present in household. * p<0.1; ** p<0.05; *** p<0.01.

Table 5. Remedial Tutoring Impacts on Endline Math and Reading Test Scores

	Dependent Variable is Test Score Percentiles				Dependent Variable is Test Score Standard Deviations			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A. Math							
Assigned to Remedial Tutoring	0.531	-0.764	-0.798	-0.68	0.01	-0.028	-0.027	-0.023
	(1.311)	(1.108)	(1.152)	(1.182)	(0.051)	(0.040)	(0.041)	(0.043)
Baseline Math Score		0.434	0.402	0.401		0.426	0.394	0.393
		(0.032)***	(0.038)***	(0.035)***		(0.037)***	(0.037)***	(0.033)***
Female				0.361				0.044
				(1.141)				(0.048)
Age				-1.3				-0.052
				(0.986)				(0.039)
Control Group Mean	38.769				0.000			
	(1.654)***				(0.065)			
	B. Reading							
Assigned to Remedial Tutoring	1.724	1.033	0.905	0.959	0.047	0.046	0.038	0.038
	(1.438)	(1.436)	(1.458)	(1.497)	(0.059)	(0.058)	(0.059)	(0.060)
Baseline Reading Score		0.558	0.471	0.452		0.52	0.443	0.428
		(0.032)***	(0.028)***	(0.029)***		(0.037)***	(0.030)***	(0.031)***
Female				4.006				0.155
				(1.234)***				(0.048)***
Age				-1.869				-0.116
				(0.791)**				(0.042)***
Control Group Mean	36.698				0.000			
	(2.124)***				(0.079)			
N	1129	1129	1129	1129	1129	1129	1129	1129

Notes: Standard Errors clustered at the school level in parentheses. Table shows Math and Reading endline impact results. In columns (1)-(4) outcome variable and lagged test-score regressor are expressed in percentiles. In columns (5)-(8) outcome variable and lagged test-score regressor are expressed in standard deviation units; (1) and (4) No controls; (2) and (5) Controls for baseline same-subject score; (3) and (6) Controls for baseline scores and school fixed effects; (4) and (8) Controls for baseline scores, school fixed effects and other student socio-demographic characteristics not shown in the table including school shift, Spanish speaking, adults in household and father present in household. * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Table 6. Heterogeneity of Tutoring Impacts on Science Achievement by Gender

	Dependent Variable is Test Score Percentiles				Dependent Variable is Test Score Standard Deviations			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	4.045	1.935	2.137	2.016	0.191	0.115	0.127	0.118
	(2.163)*	(2.080)	(2.143)	(2.117)	(0.084)**	(0.081)	(0.084)	(0.081)
Assigned to Remedial Tutoring	5.697	5.45	5.154	5.44	0.223	0.226	0.213	0.222
	(1.409)***	(1.361)***	(1.357)***	(1.430)***	(0.067)***	(0.066)***	(0.067)***	(0.069)***
Female * Assigned	-4.505	-4.089	-4.213	-4.635	-0.207	-0.196	-0.206	-0.218
	(2.565)*	(2.496)	(2.547)	(2.489)*	(0.105)*	(0.105)*	(0.108)*	(0.106)**
Baseline Score		0.641	0.556	0.551		0.448	0.389	0.384
		(0.073)***	(0.053)***	(0.054)***		(0.050)***	(0.037)***	(0.035)***
Age				-1.895				-0.089
				(0.860)**				(0.032)***
Control Group Mean	32.883				-0.084			
	(2.508)***				(0.101)			
N	1,129	1,129	1,129	1,129	1,129	1,129	1,129	1,129

Notes: Standard Errors clustered at the school level in parentheses. Table shows heterogeneity in science endline impact results by gender. In columns (1)-(4) outcome variable and lagged test-score regressor are expressed in percentiles. In columns (5)-(8) outcome variable and lagged test-score regressor are expressed in standard deviation units. (1) and (4) No controls; (2) and (5) Controls for baseline science score; (3) and (6) Controls for baseline science scores and school fixed effects; (4) and (8) Controls for baseline scores, school fixed effects and other student socio-demographic characteristics not shown in the table including school shift, Spanish speaking, adults in household and father present in household. * p<0.1; ** p<0.05; *** p<0.01.

Table 7. Heterogeneity of Tutoring Intensity by Gender

	(1)	(2)	(3)	(4)
A. Number of Tutoring Sessions Attended				
Female	-0.006	-0.009	-0.008	-0.049
	(0.021)	(0.027)	(0.110)	(0.108)
Assigned to Remedial Tutoring	4.741	4.741	4.734	4.736
	(0.464)***	(0.464)***	(0.479)***	(0.471)***
Female * Assigned	0.000	-0.002	0.038	0.052
	(0.264)	(0.262)	(0.276)	(0.268)
Baseline Score (Percentile)		0.001	-0.003	-0.004
		(0.009)	(0.004)	(0.004)
Control Group Mean	0.039			
	(0.017)**			
B. Percent of Tutoring Sessions Attended				
Female	-0.047	-0.094	0.075	-0.234
	(0.173)	(0.224)	(0.808)	(0.788)
Assigned to Remedial Tutoring	38.702	38.703	38.697	38.722
	(3.659)***	(3.663)***	(3.801)***	(3.738)***
Female * Assigned	0.322	0.305	0.517	0.599
	(2.100)	(2.085)	(2.168)	(2.101)
Baseline Score (Percentile)		0.018	-0.025	-0.03
		(0.076)	(0.033)	(0.033)
Control Group Mean	0.321			
	(0.138)**			
C. Additional Tutoring Time (Minutes)				
Female	-1.55	-1.974	3.676	-0.355
	(2.581)	(3.230)	(10.958)	(10.969)
Assigned to Remedial Tutoring	428.86	428.875	430.911	431.445
	(48.546)***	(48.590)***	(50.151)***	(49.343)***
Female * Assigned	3.417	3.261	4.277	5.045
	(25.218)	(25.028)	(26.514)	(26.400)
Baseline Score (Percentile)		0.161	-0.52	-0.606
		(0.965)	(0.402)	(0.410)
Control Group Mean	3.935			
	(1.972)*			
N	1,219	1,219	1,219	1,219

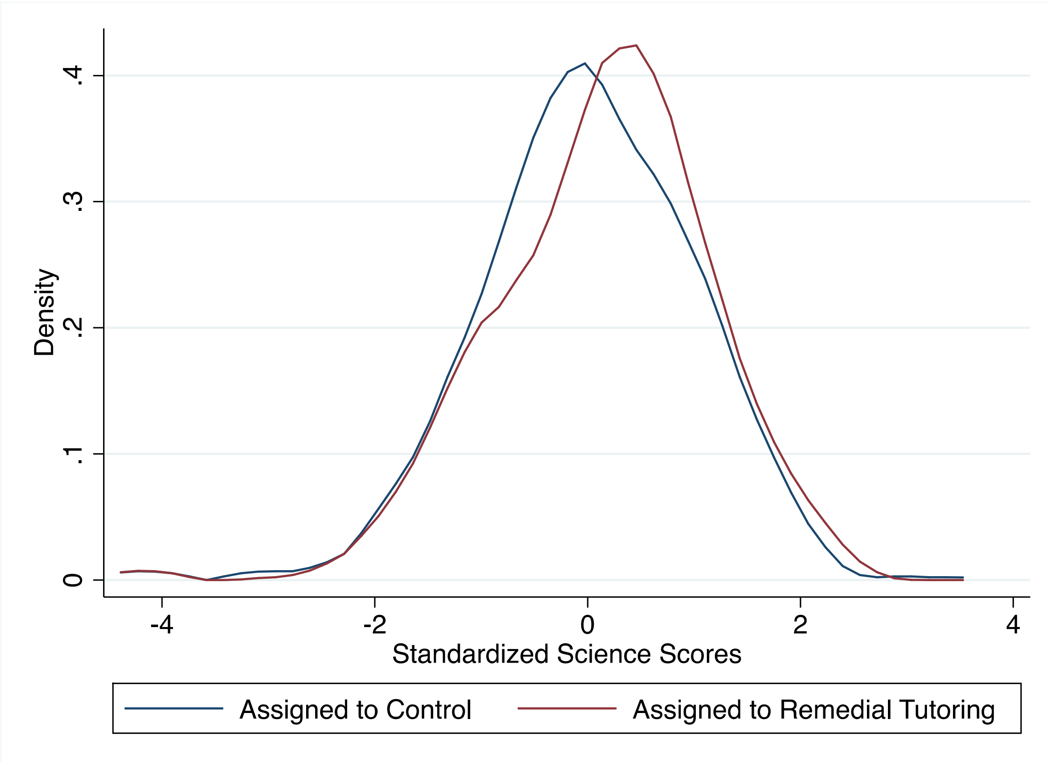
Notes: Standard Errors clustered at the school level in parentheses. Table shows heterogeneity in tutoring intensity. (1) No controls; (2) Controls for baseline science scores (percentile); (3) Controls for baseline science scores and school fixed effects; (4) Controls for baseline scores, school fixed effects and other student socio-demographic characteristics not shown in the table including age, school shift, Spanish speaking, adults in household and father present in household. * p<0.1; ** p<0.05; *** p<0.01.

Table 8. Spillovers of tutoring on science endline performance

	Dependent Variable is Test Score Standard Deviations				
	(1)	(2)	(3)	(4)	(5)
Section Fraction Assigned to Tutoring	-1.287	0.327	-0.074	-0.022	-0.112
	(0.382)**	(0.396)	(0.412)	(0.431)	(0.363)
Section Average Baseline Science Scores		0.18	0.042	0.043	0.045
		(0.070)*	(0.101)	(0.095)	(0.096)
Baseline Science Score		0.502	0.501	0.497	0.497
		(0.021)**	(0.021)**	(0.021)**	(0.021)**
Female			-0.078	-0.075	-0.119
			(0.019)**	(0.019)**	(0.120)
Age			-0.061	-0.064	-0.075
			(0.034)	(0.034)	(0.019)**
Section Fraction Assigned to Tutoring * Female					0.219
					(0.491)
N	2,246	2,246	2,246	2,246	2,246

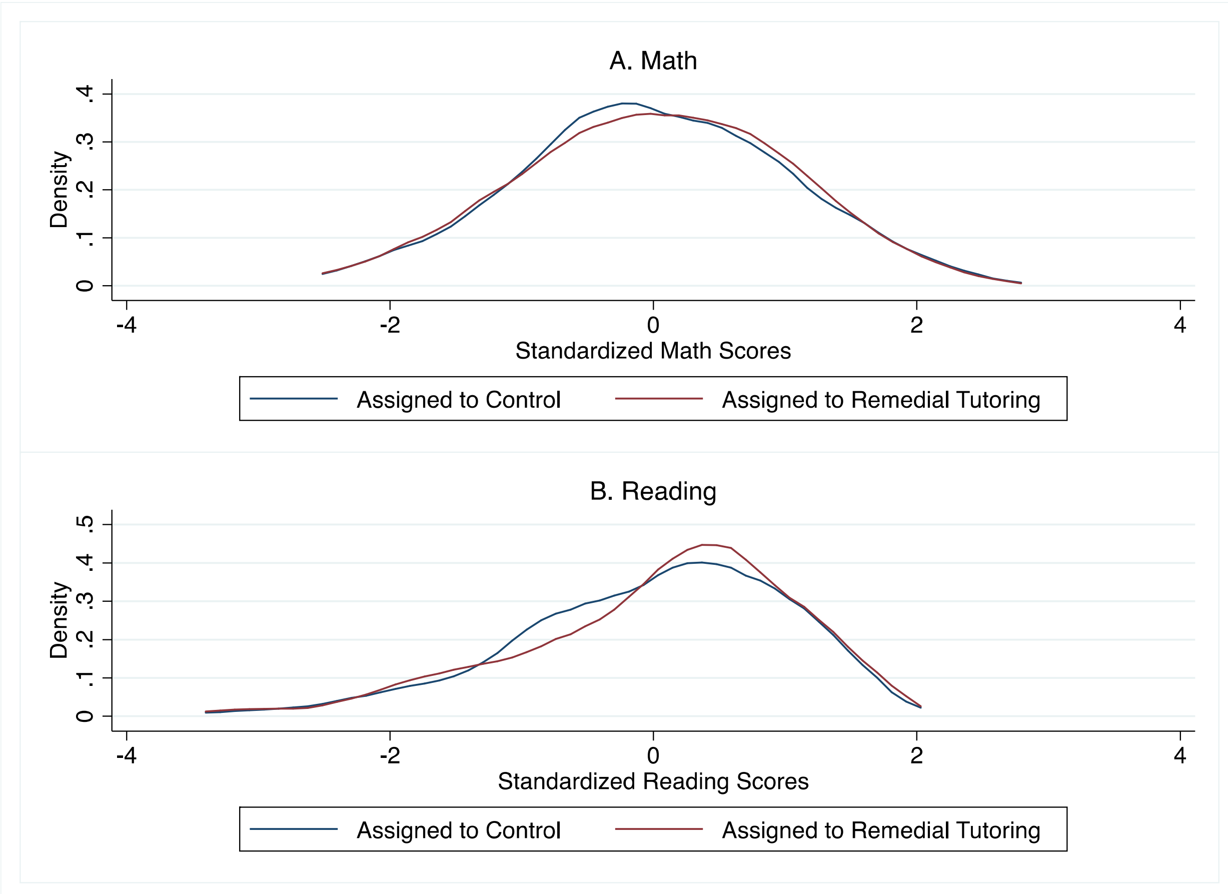
Notes: Standard Errors clustered at the school level in parentheses. Table shows spillover effects of tutoring on endline science scores. Regression results in columns (3) also include school fixed effects in addition to the reported coefficients. Regression results in columns (4) and (5) include school fixed effects and controls for school shift, Spanish speaking, adults in household and father present in household in addition to the reported coefficients. * *p*<0.1; ** *p*<0.05; *** *p*<0.01.

Figure 1. Distribution of Endline Science Test Score Impacts by Treatment Assignment Status



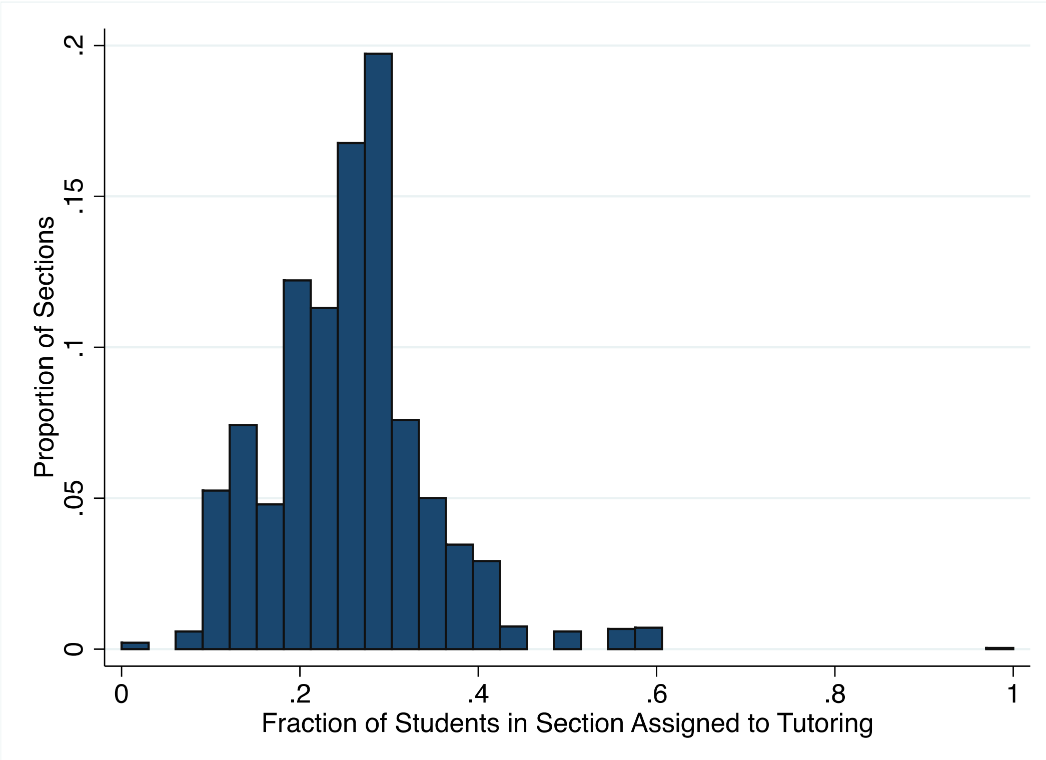
Notes: Figure shows kernel density plots of endline Science scores in standard deviation units for students assigned to tutoring treatment and control. Sample is the same as that for Table 4.

Figure 2. Distribution of Endline Math and Reading Test Score Impacts by Treatment Assignment Status



Notes: Figure shows kernel density plots of endline Math and Reading scores in standard deviation units for students assigned to tutoring treatment and control. Sample is the same as that for Table 5.

Figure 3. Variation in the fraction of students in a section assigned to remedial tutoring



Notes: Figure shows the fraction of students assigned to remedial science tutoring in the third-grade sections of the 48 schools in the evaluation sample.