NBER WORKING PAPER SERIES

TARGETED REMEDIAL EDUCATION: EXPERIMENTAL EVIDENCE FROM PERU

Juan Saavedra
Emma Näslund-Hadley
Mariana Alfonso

Working Paper 23050
http://www.nber.org/papers/w23050

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2017

Saavedra (corresponding author): University of Southern California Dornsife Center for Economic and Social Research, and National Bureau of Economic Research, juansaav@usc.edu; Näslund-Hadley: Inter-American Development Bank; Alfonso: Inter-American Development Bank. We thank Richard Murnane and Anna Saavedra for helpful comments. We also 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 USC and IPA institutional review boards. This field trial was pre-registered with the American Economic Association's RCT Registry Number AEARCTR-0000379, www.socialscienceregistry.org/trials/379. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research or the Inter-American Development Bank (IDB).

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2017 by Juan Saavedra, Emma Näslund-Hadley, and Mariana Alfonso. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Targeted Remedial Education: Experimental Evidence from Peru
Juan Saavedra, Emma Näslund-Hadley, and Mariana Alfonso
NBER Working Paper No. 23050
January 2017
JEL No. I21,I25,O15,O54

ABSTRACT

An outstanding challenge in education is improving learning among low-achieving students. We present results from the first randomized experiment of an inquiry-based remedial science-education program for low-performing elementary students in the setting of a developing country. At 48 low-income public elementary schools in Lima, Peru and surrounding areas, third-grade students scoring in the bottom half of their science classes were selected at random to receive up to 16 remedial sessions of 90 minutes each during the school year. Control-group compliance with assignment (no extra tutoring) was close to perfect. Treatment-group compliance was roughly 40 percent, or five to six remedial sessions—a 4 to 5 percent increase in total science instruction time over the school year. Despite the low-intensity treatment, students assigned to the remedial sessions scored 0.12 standard deviations higher on a science endline test. But all improvements were concentrated among boys, for whom gains were 0.22 standard deviations. Remedial education does not produce within-student spillovers to math, or spillovers on other students.

Juan Saavedra
Dornsife Center for Economic and Social Research
University of Southern California
635 Downey Way
Los Angeles, CA 90089
and NBER
juansaav@usc.edu

Emma Näslund-Hadley
Inter American Development Bank
1300 New York Avenue N.W.
Washington, D.C. 20577
EMMAN@iadb.org

Mariana Alfonso
Inter American Development Bank
1300 New York Avenue, N.W.
Washington, D.C. 20577
marianaa@iadb.org

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/379

1. Introduction

An outstanding challenge in education is improving learning among low-achieving students. Many developing countries—including most in Latin America—have centralized educational 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 also may fail to address all students' needs by teaching the same content at different skill levels (Duflo, *et al*. 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, *et al*. 2009). In Peru, for example, only 27 percent of grade-2 students meet the grade-specific proficiency level in math (MINEDU 2015). In comparative international assessments, Peruvian students typically score at the bottom. Of 65 nations participating in the 2012 application of the PISA test, Peru ranked last in both in math and natural science. Seventy-five percent of Peruvian students were low achievers in math as compared to 23 percent of OECD students (OECD 2013). In the 2013 TERCE regional study, close to 40 percent of sixth-grade Peruvian students scored at the lowest level of achievement in science (LLECE, 2015).

To address these dismal results, the government of Peru piloted two programs, in 2010 and 2012, to improve science performance in third grade, which is when students first receive dedicated science instruction. These programs, based on the 2008 national curricular standards for science, mostly consisted of training teachers to use student-centered methodologies. Consistent with the idea of many students falling behind the national curriculum, a randomized

evaluation concluded that the 2010 pilot program only improved the achievement of third-grade students who already had above-average baseline performances (Beuermann, *et al.* 2013).[1]

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. However, evidence on remedial education mostly is limited to improving basic math and literacy skills.[2] The evidence on such education suggests 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). But 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 Lima, Peru and surrounding areas—those same students left

[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)

[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 *et al.* 2000). Another US meta-analysis of 35 math 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 math (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 math and reading tutoring programs concludes that programs of moderate duration (45–85 hours) have greater effect on both reading and math 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).

behind by the previous pilot programs to improve science achievement among third graders. To our knowledge, this is the first rigorous study documenting impacts of an inquiry-based remedial science-education program targeted at early grades.

Students scoring in the bottom half of their classes on a science test administered in 2014 were assigned at random to receive throughout the school year up to 16 science remedial sessions of 90 minutes each, or to not receive any remedial support (control group). Remedial sessions followed an inquiry-based format and took place in schools—typically in the afternoon—in groups of nine students, on average. Tutors were public-sector elementary school teachers selected among volunteer candidates. Prior to the start of the program, selected tutors received content knowledge and pedagogical training, as well as detailed and highly structured materials that included flipcharts with activities for each session and formative evaluation rubrics.

Control-group compliance with assignment was close to perfect. On the other hand, treatment-group compliance was roughly 40 percent, equivalent to 5 to 6 remedial sessions—a 4 to 5 percent increase in total science instruction time over the school year. Despite the very low intensity of treatment, students assigned to remedial tutoring scored 0.12 SD higher on a science endline test. These results are striking because they were accomplished among a group of students whose achievement was not improved at all by prior universal interventions aimed at raising the quality of teaching and instruction. However, while the benefits of the remedial science program were widespread through the baseline achievement distribution of low-performing students, the program was ineffective at improving the achievement of the lowest 10 percent of students.

Moreover, all the achievement gains were concentrated among boys, for whom gains were 0.22 SD. The differing improvements between boys and girls are not explained by gender differences in treatment compliance. Our observation of remedial sessions indicates, however, that the concentration of advances entirely among boys have may stemmed from the preferential treatment given them by tutors, with whom we observed they engaged more proactively than with girls—even though the overwhelming majority of tutors in our study were women. While we cannot rule out that boys behaved more proactively than girls in these small-group tutorials, 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 (Fenema, *et al*. 1990; Mendick 2006). Our findings suggest gender gaps in science arise early, and leave open the possibility that student gender-grouping may be more relevant than instructor-student gender-matching in order to help overcome STEM achievement gender gaps (*cf.* Carrell, Page and West 2010).

We found no evidence of spillovers on science achievement onto students from the same classrooms who did not receive the remedial education. We also found no evidence that inquiry-based remedial science tutoring had within-student spillovers on achievement in math. However, we could not rule out small within-student spillovers on reading achievement.

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 designed and evaluated; 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 concluding sixth section discusses the results' limitations and possible implications.

2. Background and Program Description

In this section, we describe classroom practices and recent efforts to boost science skills in Peru that motivated the present study (subsection 2.1), and the program we evaluated (subsection 2.2).

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

Peruvian students have poor overall performance in international assessments. In a recent application of the PISA test, for example, among the 65 participating nations Peru ranked last in math and natural science (OECD 2013). One-third of Peruvian students placed in the lowest proficiency level of science, which means that they did not master even the most basic skills. The PISA assessment indicated Peruvian students lacked critical reasoning skills, as well as the abilities to analyze and synthesize information, and 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. In addition, teachers overemphasize the least cognitively-demanding topics, pose learning tasks that are not cognitively challenging, rarely provide students with feedback and when they do, it is often erroneous (Cueto, *et al*. 2006). Moreover, half of math teachers nationwide cannot perform basic math calculations (Alfonso, *et al*. 2012).

In Peru, scientific learning typically follows an explicit teacher-instruction model (*e.g.,* Clark 2012; Kirschner 2006). Lectures take up most class time, while practical work is limited. To the

extent that teachers conduct practical work they do so themselves, reducing students' opportunities for hands-on learning (Loera, *et al.* 2013; Näslund-Hadley, *et al.* 2014).

To address some of the country's educational challenges in math and science, Peru's government in 2010 piloted a program aimed at promoting critical-thinking and scientific-reasoning skills in third grade, which is when students first receive dedicated scientific instruction. This program, based on the 2008 national curriculum, included fields of study such as the physical world, the human body, living beings and the environment. A key component of the pilot science program was teacher training, with a 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 the 2010 pilot only improved science scores among boys in urban areas and among students who already had above-average baseline performance (Beuermann, *et al.* 2013). In a 2012 follow-up, remedial-session working groups were separated by gender for some activities to ensure girls got hands-on experience; in addition, rural teachers received additional mentoring. While these adjustments helped close the gender and urban-rural achievement gaps, the program still had no impact among students in the bottom half of the baseline score distribution (IPA 2014). As a consequence, the science achievement gap between high and low performers widened.[4] These earlier results motivated the present study, which

[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)

investigated whether remedial sessions targeted to the lowest-performing students would reduce science achievement gaps between high and low performers.

2.2 Program Description: The targeted remedial science program

The targeted remedial science program aims to help low-performing students master scientific theoretical and practical knowledge through inquiry-based methods. The goal is for students, when confronted with an unfamiliar situation, to develop relevant answers through critical thinking and collaborative work. As a byproduct, the program seeks to promote healthy study habits, academic motivation and a love of learning.

Universidad Cayetano Heredia in Lima developed the program's structure and contents. The program has four components: a) development of pedagogical materials, b) selection and training of tutors, c) selection of students, and d) implementation of remedial sessions in schools.

a. Pedagogical materials

Universidad Cayetano Heredia employed two local pedagogy specialists to develop materials for the remedial program: a specialist in primary education and another in science education. The materials' content is based on the 2008 national curricular standards for teaching science to third-graders.

To bridge gaps in tutors' knowledge, the specialists developed detailed and highly structured materials that included flipcharts with activities for each session and formative evaluation rubrics. That is, the materials combined elements of explicit instruction with inquiry-based activities. In this approach, remedial sessions began with a challenge/question. For example, as part of a weather module, students explored why Lima often is shrouded 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 then were 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.

b. Selection and training of tutors

Universidad Cayetano Heredia selected 16 tutors for the program – 15 of whom were women, just as are most public school teachers in Peru. Universidad Cayetano Heredia chose a male tutor for students in schools located in high-crime areas. Tutor selection took place between March and May 2014. Selection criteria included: i) at least two years experience teaching primary school, ii) a positive attitude towards the teaching and learning of science, iii) assertive communication and class-management skills, and iv) the ability to create respectful, empathetic and tolerant relationships with children.

Tutors were local primary or secondary public-school teachers, although not necessarily in the schools where they provided remedial science sessions. Tutors were paid a wage of US$10 per hour, including transportation – less than what primary education teachers earn on average (US$14 per hour). Tutors were assigned to participating schools based on geographic proximity to their residences.

Once selected, tutors participated in a Universidad Cayetano Heredia-organized training workshop led by the two education specialists. The workshop took place before the start of the 2014 school year and lasted 20 hours, split over six days. The 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 in applying these foundational principles to the tutoring activities to best engage students. Some of these approaches include encouraging and discussing different points of view, sequencing contents to follow the children's logic and viewing unfamiliar situations from new perspectives. In the workshop, the specialists and tutors also reviewed each of the 16 remedial science sessions' principles, pedagogical approaches, content and activities. Finally, throughout the school year during which remedial sessions took place, the two specialists provided ongoing support to tutors.

c. Selection of students

The program targeted low-performing third-grade science students in 48 public elementary schools in Lima and surrounding areas (sample selection details below). Baseline performance was assessed through a written test administered during class in May 2014 (two months after the start of the school year, which begins in March). Within each school, the program targeted the bottom half of scorers. Eligibility for participation then was determined using a lottery (details below).

d. Implementation of remedial sessions

Remedial sessions took place in each of the 48 participating schools, in a total of 70 groups of students. There were more groups than schools because some of the schools have large third-grade classes or more than one third-grade section. Each tutor was assigned, on average, to five tutoring groups (tutors had as few as three and as many as seven). Mean group size was nine students.

The 16 weekly, 90-minute remedial sessions began in July 2014, four months after the start of the school year, and finished in November with the school year. In total, students in the program had the opportunity to receive 24 hours of remedial education, a 14 percent increase in instructional time relative to the in-class science lessons. Based on evidence from math and reading tutoring programs, programs of moderate duration have greater impact on achievement than programs of longer duration (Lauer, *et al.* 2006).

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, for students attending school in the afternoon, sessions were scheduled in the morning or on Saturday. In the first session, students received a workbook, Making and Learning Science, which describes various scientific inquiry activities students could pursue independently.

Each tutor was responsible for coordinating and scheduling sessions with each group. Tutors approached school principals and third-grade teachers to explain program details and seek support in promoting attendance of eligible students. Tutors also were responsible for inviting parents of eligible students to information sessions explaining the goals of the remedial science program, the approach and expected benefits.

Parental attendance to the information sessions was low. To ensure all parents were informed about the availability of the remedial science program, students were asked to bring home an information sheet to be signed by a parent and returned. Some tutors visited students' homes to contact parents. In total, about 50 percent of the parents 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

3.1 Evaluation sample

To determine students' eligibility for the program, in May 2014 we collected test scores from third-graders in 51 Lima-area public elementary schools. Out of these schools, three were not selected for this study: Two had fewer than eight third-graders, and we did not want to risk stigmatizing eligible students; at the other, we could not contact tutoring-eligible children. Of the remaining 48 schools, 39 participated in the 2012 pilot science education program – so we choose them to facilitate access to the tutors, as the administrators and faculty members there had prior contact with the training staff from Universidad Cayetano Heredia. The remaining nine schools had been chosen at random among comparable schools in the poorest localities of the Lima area.

The typical school in the evaluation sample had two third-grade sections and 51 third-grade students. Principals of these schools had, on average, 6.3 years of experience in that position; teachers had an average of 5.6 years of experience, all but one year at their current schools. About 14 percent of the teachers participated in the 2012 pilot program. Of the students, about 55 percent were boys and the average age was just over 8 years old.

3.2 Experimental design

In May 2014, we collected baseline science test-score data from 2,399 third-grade students in the 48 schools of the evaluation sample. The test was a simplified version of the exam administered as part of the 2010 and 2012 science pilot programs implemented in Lima. That earlier test, and others covering math and Spanish, measured third-grade skills based on Peru's new (2008) basic education curriculum and national study plan. In science, the curriculum

includes the physical world, preservation of the environment, the human body and health, and animals and their environment.

Test questions address a mixture of content and critical-thinking skills. Content questions ask, for example, how different food groups can help us stay healthy, and the identification of Peruvian animals. As an example of a critical-thinking question, students are asked why a snow cone turned into red water when a little girl left it on a bench while playing. Students individually in writing complete the tests, administered and timed by a supervisor.

The bottom half of scorers in the baseline test within each school were targeted for the remedial science program—1,219 students in total. Among them, we randomly assigned participation eligibility, stratifying by school and gender. In practice, we only conducted 95 lotteries ([48 x 2] – 1) because in one school only boys scored in the bottom half of the test-score distribution. In the final evaluation sample, we assigned to treatment 609 students (331 boys and 278 girls). Meanwhile, the other 610 students (337 boys and 273 girls) were assigned to control conditions—that is, they did not receive remedial science education.

3.3 Randomization balance

Randomization balanced characteristics across students assigned to treatment and control (Table 1). These characteristics included gender, age, school shift (morning or afternoon), whether the student spoke Spanish, the number of adults in the household, whether the father of the student was present in the household and baseline test scores (Panel A, Table 1). At baseline, boys and girls scored 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 was math scores among boys (Panel B, Table 1).

4. Data and Empirical Strategy

4.1 Data

We employed three data sources, the first being the baseline test and socio-demographic questionnaire collected from third-grade students in the sample's 48 schools.[5] The second source was data on student attendance to the tutoring sessions, collected by the tutors (*i.e.*, compliance with treatment assignment). These were collected for 12 out of 16 tutoring sessions; for the first four sessions, tutors did not take attendance. We measured student attendance three different ways: i) number of tutoring sessions attended, ii) fraction of tutoring sessions attended and iii) total additional minutes spent in tutoring. The final data source was endline test and student survey data, collected at the end of the school year in November 2014—about five months after the start of the remedial sessions.

Endline attrition was low, uncorrelated with either treatment status or the demographic composition of randomization groups (Table 2). More than 90 percent of students assigned to control took the endline test (bottom of Column 1, Table 2). Students assigned to treatment were 1 percentage point more likely to take the test – but the difference was 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 also was balanced, as the F-statistic of the joint test of interactions between baseline characteristics and treatment assignment is 1.50 (p-value= 0.19).

4.2 Empirical strategy

[5] The questionnaire was designed and validated for use among elementary students. For example, it did not include questions about income, but rather about dwelling characteristics and assets.

In all tables, we begin by showing unadjusted mean differences in outcomes between students assigned to remedial science 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 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 sequentially added the control variables to document the stability of estimates of τ to alternative sets of control variables. We also explored heterogeneity by gender in the estimates of τ.

5. Results

We discuss results on compliance with treatment assignment (subsection 5.1), endline science achievement (subsection 5.2), endline math and reading achievement (subsection 5.3), heterogeneity by gender (subsection 5.4) and spillovers to non-treated students in the treatment schools (subsection 5.5).

5.1 Compliance with treatment assignment: Attendance to remedial science education sessions

Compliance with treatment assignment among students assigned to control was very high, so control group contamination was negligible. On average, students assigned to control conditions

attended 0.04 tutoring sessions (Panel A, Table 3), receiving three additional minutes of tutoring time (Panel C, Table 3).

However, compliance among students assigned to the remedial treatment was low. In the 12 remedial sessions for which we collected attendance data, the average student assigned to treatment attended 4.7 sessions (Panel A, Table 3). This represents a compliance rate among treatment-eligible students of 39 percent (Panel B, Table 3). Students assigned to remedial sessions received, on average, 430 minutes of additional science education relative to total instruction time over the school year (Panel C, Table 3). This implies that rather than the intended additional 14 percent of science instruction time, the average student received an additional 4 percent of science instruction time.

Since we did not take attendance in the first four remedial sessions, these estimates represent a lower bound on actual attendance rates. Assuming similar attendance in the first four sessions as in the 12 remedial sessions observed, the average student assigned to treatment likely attended 6.3 remedial sessions, receiving 573 additional minutes of tutoring – a 5.3 percent increase in science instructional time over the school year.[6]

Baseline science scores and students' gender were uncorrelated with tutoring attendance (Columns 2-4, Table 3). The low attendance was the result of a combination of factors, including failure to effectively promote the program and its benefits. In addition, students may have had time conflicts with other responsibilities, as 43 percent of Peru's 5 to 17 year olds are economically active, generally combining school with work. Although child labor is 40 percent more prevalent in rural areas, children in urban areas also are economically active, mainly as

[6] One can obviously construct a different upper bound by assuming all students assigned to treatment attended all initial four sessions. This upper bound is unrealistic, however, given that only 50 percent of parents knew about the availability of the remedial education program for their children.

street vendors (ILO, 2009). The prevalence of child labor increases with age, which would be consistent with our finding that younger students were more likely to attend remedial sessions (Column 4, Table 3). Moreover, in the afternoon or during weekends, children might have been helping at home by taking care of younger siblings while their parents were working.

5.2 Endline science test scores

Remedial tutoring increased endline science scores. When measured in percentiles of the test score distribution, the estimate of treatment's impact is between 3 and 4 percentiles. The ITT estimates were 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 SD (Columns 5-8, Table 4).

ITT impact estimates of remedial tutoring were substantial in magnitude once we accounted for the intensity of treatment among students assigned to tutoring. For example, the 39 percent difference in compliance rates between students assigned to tutoring or control (Panel B, Table 3) would imply estimates for the treatment on the treated of 0.30-0.36 SD. We could not, however, give a causal interpretation to this Wald estimate because doing so would require assuming that tutoring only affected student outcomes through participation in remedial sessions. This assumption may be challenged if, for example, non-participants benefited indirectly through improved regular classroom learning as a result of fewer underperforming students delaying the pace of learning. We explore the empirical support for these potential spillover effects in subsection 5.5.

The gains of tutoring on endline science achievement accrued to students who scored at baseline between negative one standard deviation and average (Figure 1). In the sample, this corresponded roughly to students at or above the 10th percentile. This result suggests that while

the benefits of remedial science tutoring were widespread among low-performing students, the program was ineffective at improving the achievement of the lowest 10 percent of students.

5.3 Endline test scores on other subjects: Math and Reading

Remedial tutoring in science appeared to shift the distribution of math and, particularly, reading scores for students in the middle of the low-scorers distribution (Figure 2). However, point estimates of the effect of remedial tutoring on endline math achievement tests were 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 were positive, although small. Nevertheless, we cannot reject the hypothesis 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 targeted remedial science program has portability to other literacy skills, the effect may be small.

5.4 Heterogeneity by gender

The effects of the targeted remedial science program on endline science achievement were entirely driven by gains among boys. Tutoring assignment increased science scores for boys by about 5 percentiles (Columns 1-4, Table 6), about 0.22 SD (Columns 5-8, Table 6). For girls, tutoring impacts were negligible. Coefficients on the interaction term were about -4.5 percentiles or -0.21 SD. Estimates of coefficient on the interaction term generally were 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 gender gap is differences in treatment intensity (compliance) between boys and girls. We did not find empirical support for this conjecture. Boys and girls

were equally likely to attend tutoring sessions; the coefficients on the interaction term were always small, relative to the main treatment effect, and never statistically significant (Table 7).

Our observation of remedial sessions indicates, however, that the concentration of gains entirely among boys may have stemmed from the preferential treatment given boys by tutors, who we saw engaging more proactively with boys than girls—even though the overwhelming majority of tutors in our study were women. We cannot rule out, however, that boys behaved more proactively than girls in these small-group tutorials.

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 (Fenema, *et al*. 1990; Mendick 2006)

5.5 Spillovers

Because we randomized at the student level within schools, it was possible that tutoring generated spillovers on students who were ineligible to participate. For example, non-participants may have benefited indirectly through improved regular classroom learning because there were fewer underperforming students to delay the pace.

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

Figure 3 shows this variation was considerable. In some sections, no students received remedial tutoring, while there were several sections in which anywhere between 20 and 60 percent of students were assigned to tutoring. In one section, all students were assigned.

We took advantage of this variation to estimate learning spillovers of tutoring onto other students. Specifically, under a linear-in-means peer-effects model, if (positive) spillovers existed, student achievement should have been be higher in sections with a higher fraction of students assigned to remedial tutoring. To test this hypothesis, we used the following regression model:

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

where $\overline{D_c}$ is the section-level fraction of students eligible for remedial tutoring and $\bar{Y}_{sc,t-1}$ is the same-subject average baseline performance of section c in school s. In this specification, we included baseline section-average performance; without it, $\overline{D_c}$ is mechanically (negatively) correlated with the outcome variable. This is so because tutoring targeted low-performing students, meaning 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 had in mind asked whether among two sections of the same school with similar baseline composition, students in the section with a higher number of students receiving tutoring would perform better than students in the section with fewer tutoring-assigned students. As before, we also allowed error terms to arbitrarily co-vary within schools.

We found no evidence of tutoring-related learning spillovers among other students in the same section. Without controls, the fraction of students assigned to tutoring in a section negatively correlated with endline science achievement (Column 1, Table 8). Including the section average and baseline individual achievement flipped 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 included school-fixed effects and full demographic controls, the coefficient on the fraction of students assigned to tutoring in a section was small, negative and not statistically significant (Columns 3 and 4, Table 8).

6. Conclusion

Many developing countries—including most in Latin America—have centralized educational systems with standardized curricula and textbooks, but struggling students often fall behind the national curriculum at an early age. The evidence to date on remedial math and literacy education suggests direct instruction may be an effective pedagogical model to improve the performance of low-achieving students in these subjects. However, research on whole-class science instruction suggests that inquiry-based teaching—in which students engage in hands-on practical work with different degrees of instructor guidance—improves learning more than traditional classroom practices. It is unclear whether inquiry-based instruction is an effective educational approach for remedial education targeting low-achieving, early-grade students.

Our study is the first that rigorously evaluates an inquiry-based remedial science education program targeted at lower-performing students in early grades. Also, it is the first randomized experiment of a science-tutoring program for small groups of lower-performing students in Latin America.

Our experimental results suggest struggling students' achievement can be improved through low-intensity, inquiry-based targeted remedial science education. The benefits of the remedial science program were widespread through the baseline achievement distribution of low-performing students, although the program was ineffective at improving the achievement of the lowest 10 percent of students. These results are striking because they were accomplished among

a group of students whose achievement was not improved at all by prior universal interventions aimed at raising the quality of teaching and instruction.

The findings suggest low-performing students can learn through inquiry-based pedagogical approaches. The targeted remedial science education model could easily be expanded to provide intensive academic support at a large scale for students who fall behind. The project would be straightforward to replicate because the tutors are local and the training is short.

However, we identified two important qualifiers to keep in mind when thinking about bringing this remedial science education model to scale. First, the inquiry-based remedial science-education program did not significantly improve achievement among girls. A challenge remains to identify instructional models that help mitigate instructor and student stereotypes and attitudes with regards to gender and science. Second, the overall effectiveness of the remedial education model was achieved despite a very low compliance rate. The effect could potentially be increased with greater compliance, and the compliance rate may improve by more clearly disseminating the program and promoting its benefits among parents and students. Because many students are either economically active or providing help at home by taking care of younger siblings, what may help improve attendance is a more flexible schedule or the provision of snacks to facilitate an extended school day.

References

Alfonso, Mariana, *et al*. 2012. "Panorama General de la Educación en América Latina y el Caribe." *Educación para la Transformación*, Marcelo Cabrol y Miguel Szekely (Eds.). IDB. Washington, DC.

Banerjee Abhijit, *et al*. 2007. "Remedying education: evidence from two randomized experiments in India." *Quarterly Journal of Economics,* 122:1235–64

Beuermann, Diether, *et al*. 2013. "The Pedagogy of Science and Environment: Experimental evidence from Peru." *The Journal of Development Studies*. 49(5): 719-736.

Brickman, Peggy, *et al*. 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., *et al*. 2010. "Sex and Science: How professor Gender Perpetuates the Gender Gap." *The Quarterly Journal of Economics*, 125(3): 1101-1144.

Clark, Richard E., *et al*. 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, *et al*. 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, *et al*. 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, *et al*. 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, *et al*. 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, *et al*. 1990. “Teachers' attributions and beliefs about girls, boys, and mathematics.” *Educational Studies in Mathematics*, 21: 55-69.

Furtak, Erin M., *et al*. 2012. “Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis.” *Review of Educational Research*, 82 (3), 300–329.

Glewwe, Paul, *et al*. 2009. “Many Children Left Behind? Textbooks and Test Scores in Kenya.” *American Economic Journal: Applied Economics*, 1(1): 112-35.

Harris, Christopher J., *et al*. (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., *et al*. 1989. "The Effects of Direct Instruction with Rural Remedial Students: A brief report." *Reading Improvement* (26) 88-93.

Kirschner, Paul. A., *et al*. 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., *et al*. 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, *et al*. (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, *et al*. 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., *et al*. 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., *et al*. 2011. “Effective programs for struggling readers: A best evidence synthesis.” *Educational Research Review*, 6(1): 1-26.

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 scored in the bottom half of the baseline science test administered in May 2014 in 48 Lima-area public elementary schools.

Table 2. Endline Attrition Analysis

	Dependent Variable is 1 if Student Took Endline Test; 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 remedial science tutoring and to control conditions. Sample is 1,219 third-grade students who scored in the bottom half of the baseline science test

administered in May 2014 in 48 Lima-area public elementary schools.* $p<0.1$; ** $p<0.05$; *** $p<0.01$.

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

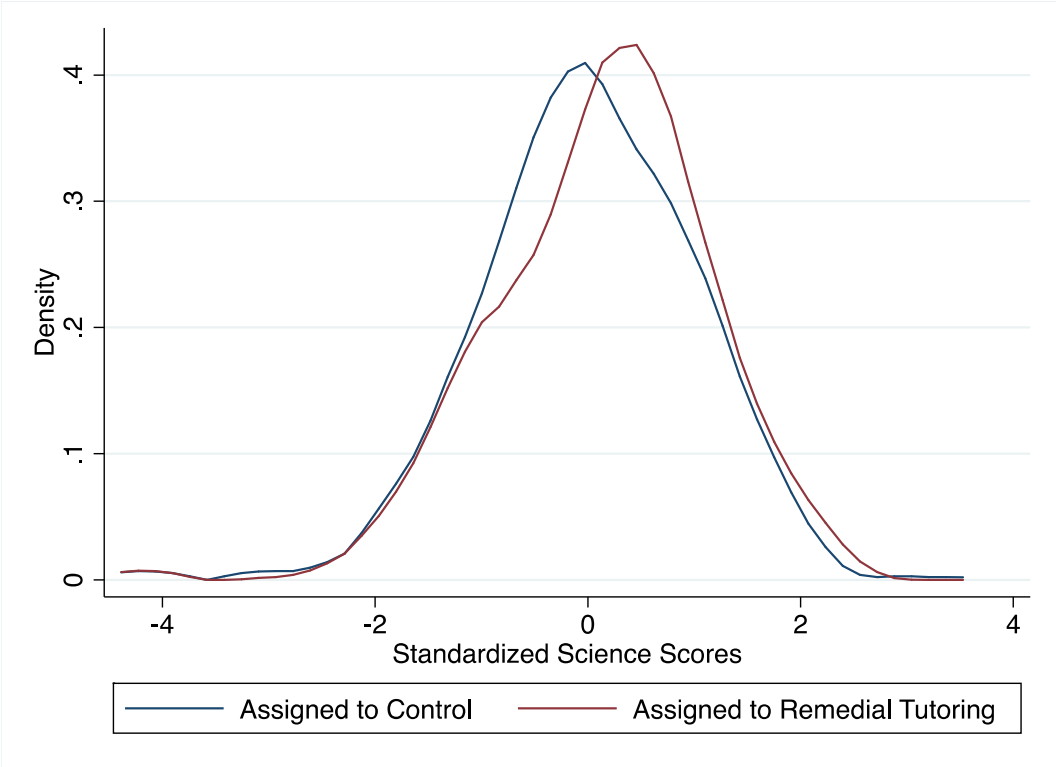
Notes: Standard Errors clustered at the school level in parentheses. Table shows heterogeneity in tutoring intensity. (1) no controls; (2) controls for baseline science scores (percentile); (3) controls for baseline science scores and school-fixed effects; (4) controls for baseline scores, school-fixed 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

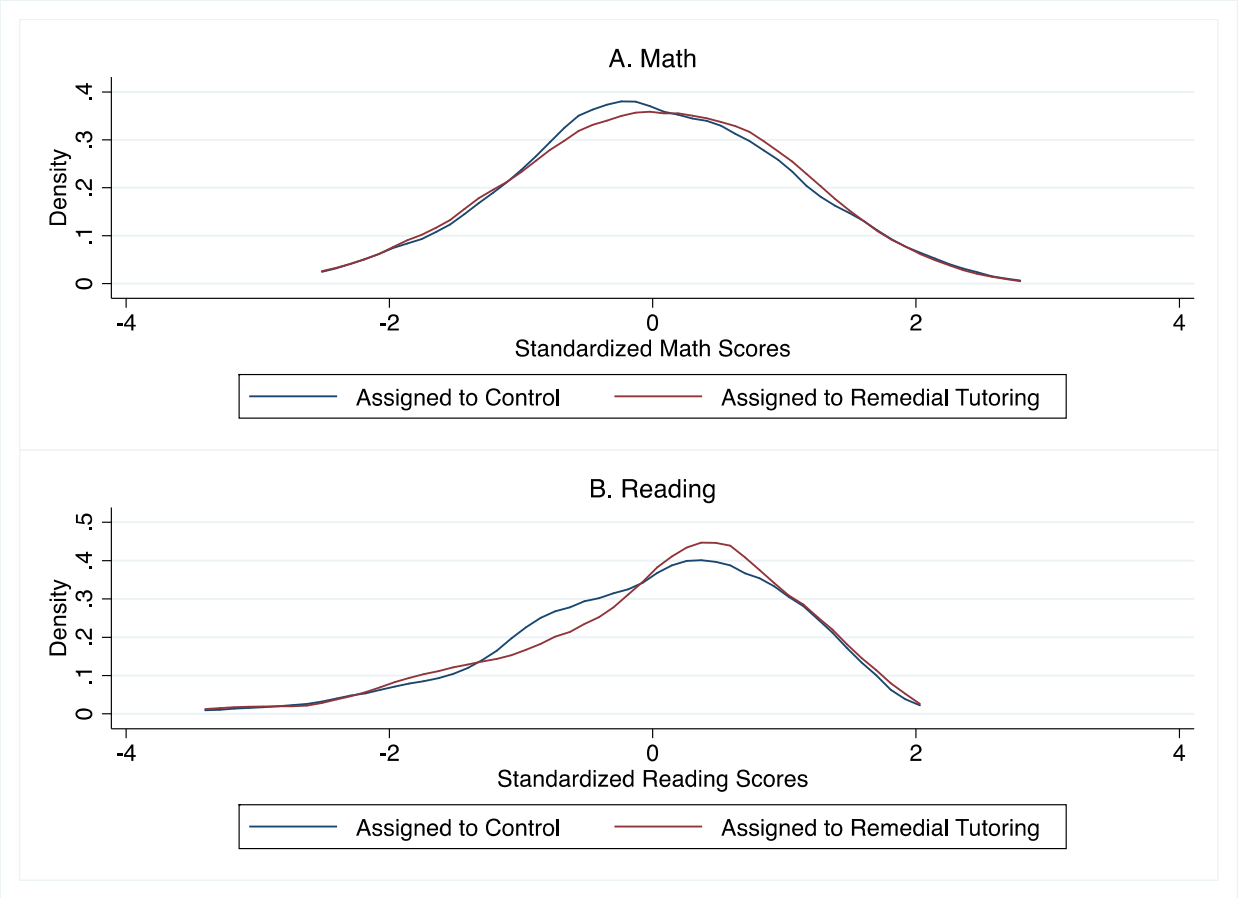
Notes: Standard Errors clustered at the school level in parentheses. Table shows spillover effects of tutoring on endline science scores. Regression results in columns (3) also include school-fixed effects in addition to the reported coefficients. Regression results in columns (4) and (5) include school-fixed 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



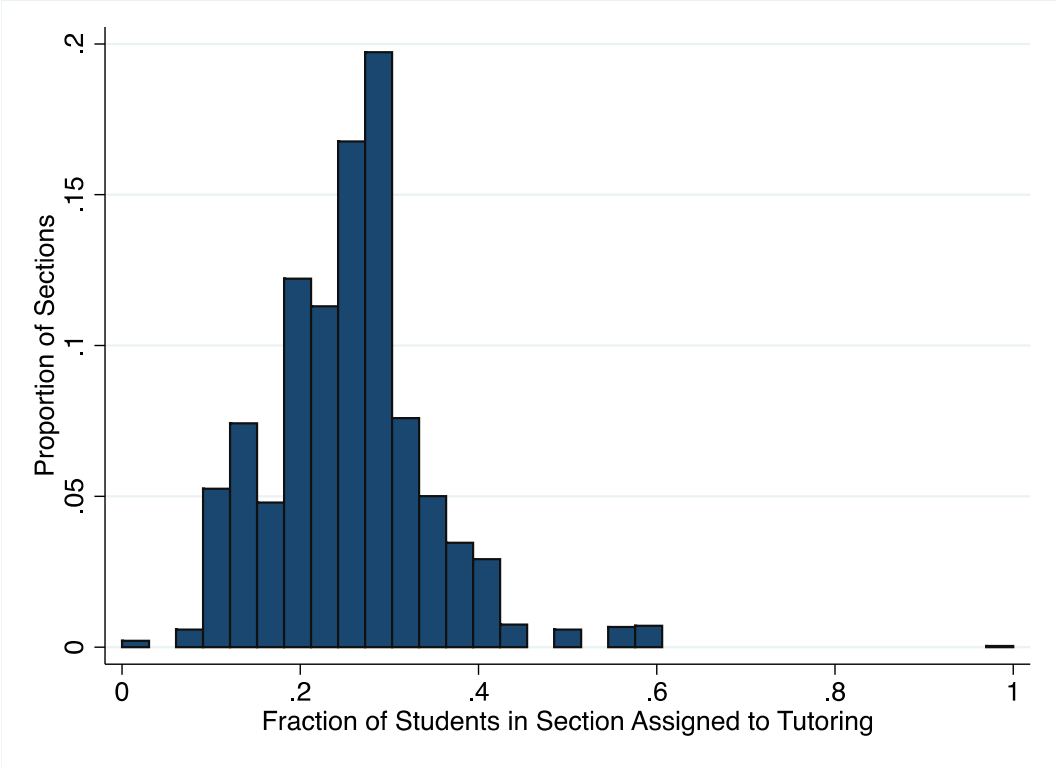
Notes: Figure shows kernel density plots of endline science scores in standard-deviation units for students assigned to tutoring treatment and control. Sample is the same as that for Table 4.

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.