Institute for Fiscal Studies

Samuel Berlinski
Matias Busso
Michele Giannola

Working paper
22/02

Helping struggling students and benefiting all: peer effects in primary education

Economic and Social Research Council

Helping Struggling Students and Benefiting All: Peer Effects in Primary Education*

Samuel Berlinski Matias Busso

Michele Giannola

January 3, 2022

Abstract

We exploit the randomized evaluation of a remedying education intervention that improved the reading skills of low-performing third grade students in Colombia, to study whether providing educational support to low-achieving students affects the academic performance of their higher-achieving classmates. We find that the test scores of non-treated children in treatment schools increased by 0.108 of a standard deviation compared to similar children in control schools. We interpret the reduced-form effect on higher-achieving students as a spillover effect within treated schools. We then estimate a linear-in-means model of peer effects, finding that a one-standard-deviation increase in peers' *contemporaneous* achievement increases individual test scores by 0.679 of a standard deviation. We rule out alternative explanations coming from a reduction in class size. We explore several mechanisms, including teachers' effort, students' misbehavior, and peer-to-peer interactions. Our findings show that policies aimed at improving the bottom of the achievement distribution have the potential to generate social-multiplier effects that benefit all.

Keywords: peer effects; remedying education.

JEL Classification: D62, I21, I25, J01

*Berlinski: Research Department, Inter-American Development Bank and IZA (email: SAMUELB@iadb.org); Busso: Research Department, Inter-American Development Bank (email: MBUSSO@iadb.org); Giannola: University of Naples Federico II, CSEF and the Institute for Fiscal Studies. (email: michele.giannola@unina.it). Corresponding author: Samuel Berlinski (1300 New York Avenue NW, Washington DC, 20010, US. Ph: +1-202-623-3842). We thank Orazio Attanasio, Imran Rasul, Michela Tincani and many seminar participants for helpful comments and discussions. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

1 Introduction

This paper studies whether providing educational support to low-achieving students affects the academic performance of their higher-achieving classmates. In the more than 50 years since the publication of the Coleman Report (Coleman (1968)), a large body of research in economics, education and sociology has documented the central role played by peers in determining academic outcomes at all education levels.[1] Of particular interest is the effect that low-achieving students can have on the performance of the rest of their classmates. Recent studies suggest that these students are detrimental to their higher-achieving peers' academic performance (e.g. Carrell & Hoekstra (2010); Lavy, Paserman, & Schlosser (2011); Imberman, Kugler, & Sacerdote (2012)). Moreover, there is evidence that these effects are persistent and translate into lower educational attainment and reduced earnings (Carrell, Hoekstra, & Kuka (2018)).

The existence of negative externalities that low achievers may have on other students provides a compelling justification that underscores why *all* parents and policy makers should be concerned about how to properly support this group of students – over and above society's wider interest in providing low achievers with the skills they need to succeed in school and in the workplace. Yet, previous work has largely been limited to describing the phenomenon rather than studying potential policies that could attenuate the impact of low-achieving peers.

In this paper, we exploit the randomized evaluation of a remedying education program that targeted struggling students within a class, to study whether an exogenous improvement of the skills of students at the bottom of the test-score distribution can generate gains for the rest of the class through achievement peer effects. The intervention we consider aimed to improve reading among low-achieving third-grade students in Colombia. At the beginning of the school year, all students were tested to determine their baseline literacy level. Students with baseline reading scores lower than a certain threshold were deemed eligible to receive the tutoring classes.[2] Schools were then randomized into treatment and control groups. In treatment schools, eligible children were taken out of the regular classes to work in small groups with a qualified tutor, who followed a structured pedagogical curriculum for 40 minutes, three times a week. In control schools, eligible children continued their classes as

[1] The Coleman Report concluded that much of the achievement gap between white and black students could be attributed to differences in the composition of peers these students faced in American public schools. For studies that analyze peer effects in elementary and secondary schools, see, for example, Hoxby (2000), Hanushek, Kain, Markman, & Rivkin (2003), and Whitmore (2005). See Sacerdote (2001), Zimmerman (2003) and Stinebrickner & Stinebrickner (2006) for evidence at the university and college levels.

[2] This eligibility threshold was determined by local pedagogues. It was based on the skill level expected from a second-grade student.

usual. The intervention improved literacy skills of low-achieving students by one-third of a standard deviation (Marinelli, Berlinski, & Busso (2021)).

The research design naturally generates two groups of students within the same class: low-achieving students who were eligible to receive the intervention, and higher-achieving students who were not eligible. (Henceforth, we refer to the students whose scores were low enough to be eligible for the tutoring as low achievers; and we refer to their classmates whose scores were above the threshold for eligibility for the tutoring as higher achievers.) Determination of students' eligibility for the tutoring program took place prior to schools' randomization into treatment and control status, allowing us to identify these two groups of children both in treated and control schools.

We find that non-eligible children in treated schools scored 0.108 of a standard deviation higher than similar children in the control group. This coefficient is sizable and represents roughly 30 percent of the treatment effect we measure on the *eligible* students. This result is economically meaningful, and its magnitude can be compared to a more commonly proposed school-level reform, tracking by prior achievement (Duflo, Dupas, & Kremer (2011)).

We interpret the reduced-form effect on higher achievers as a spillover effect within treatment units, and we estimate linear-in-means models of peer effects. Credibly identifying peer effects is challenging given the well-known issues of selection, reflection, and correlated unobservables (Manski (1993)). We overcome these identification challenges by exploiting the experimentally induced variation in the outcome of a sub-set of individuals in the peer group. This approach is defined by Moffitt (2001) as a *partial population experiment*. Randomization of the program solves the reflection problem as it induces exogenous variation in the outcomes of low-performing children without *directly* affecting higher-performing students. Second, random assignment implies that the treatment is orthogonal to all observables and unobservable characteristics, solving the problem of correlated unobservables. Finally, because peer groups are established before the policy change and remain fixed throughout the experiment, endogenous group membership is not an issue. We can think of peer effects as being conditional on any selection into groups that might have taken place prior to the experiment.

We find strong evidence of peer effects in academic outcomes. A one-standard-deviation increase in peers' *contemporaneous* test scores increases individual reading score by 0.679 of a standard deviation. We also find evidence of non-linearities, with stronger effects for students at the top of the achievement distribution.

The term peer effect is generally used as an umbrella term that comprises *any externality*, implying

that peers' outcomes have an impact on an individual's outcome. Peers can affect learning outcomes either *directly*, through peer-to-peer interactions or misbehavior, or *indirectly*, by affecting teachers' effort and practices (Sacerdote (2011)). We seek to distinguish between these alternative mechanisms because such distinctions might be key for the design of effective education policies. Thus, we provide evidence that speaks to these different mechanisms. Using survey data on teachers, we cannot reject the null hypothesis that teachers continued with classroom practices that they were using prior to the start of the intervention. At the same time, we find suggestive evidence that a reduction in classroom disruption may have driven part of the results. This suggests that low levels of achievement foster disruptive behavior, and that interventions that only affect learning without directly targeting behavior can relax the constraints posed by low-achieving students on the rest of their classmates. Finally, by exploiting heterogeneity in the impacts of tutoring on low-achieving students within treatment school, we find evidence suggesting *direct* peer-to-peer learning interactions.

One potential concern with the interpretation of our results is that by removing low-achieving peers from the classroom, higher-achieving students experienced a reduction in class size, which in turn could have had a *direct* impact on their performance. We provide three pieces of evidence against this interpretation. First, the classroom-size reduction was modest. The average tutorial size was five (and size was capped at six); there was only one tutorial operating in each school at any given time; and eligible children were randomly assigned to tutorial groups independently of the classroom they belonged to. Thus, in an average class of 31 students, the number of students decreased by just three students. Using existing estimates from the literature, we show that this class-size reduction can explain at most a tenth of the reduced-form effect on higher-achieving children. Second, the remedying tutorials did not necessarily take place during regular literacy lessons. For this reason, if the reduction in class size were behind the reduced-form effect on higher-achieving students' test scores, we would expect to see similar effects on subjects other than literacy. We do not. Third, we find homogeneous effects on higher-achievers in classes that (by virtue of the random assignment of eligible children to tutorial groups) experienced larger or smaller reductions in size.

This paper stands out from the literature on peer effects in education in a number of fundamental ways. First, in contrast with most of the previous literature, we study the impact of peers' contemporaneous achievement – the *endogenous* effect in the terminology of Manski (1993) – on individual outcomes directly, as opposed to peers' background characteristics, such as gender, race, or

prior achievement.[3] This is particularly important given that research demonstrates that once peers' achievement is property controlled for, these background characteristics do not matter for student outcomes (Hoxby & Weingarth (2005)).[4] Moreover, peer effects stemming from background characteristics do not entail a social-multiplier effect (Sacerdote (2011)). On the other hand, effects stemming from peers' contemporaneous achievement have the potential to generate social-multiplier effects.[5] In our setting, the beneficial effects of improving the academic achievement of low-achieving students spill over onto non-treated students, magnifying the total output of the program.

Second, we focus on peer effects in naturally occurring groups, and exploit the random variation in the outcomes of a subset of group members. This distances our work from that strand of the literature that uses the random allocation of students to groups. This distinction is particularly important given that opportunities to randomly assign peers are rare in real-world settings – whereas the possibility of randomly treating a subset of individuals within a group might not be so rare.[6] Moreover, a particularly important issue is whether the results in those studies that exploit the random allocation of peers are generalizable to naturally occurring peer groups. The results in Carrell, Sacerdote, & West (2013) directly speak to this issue by highlighting how exogenously manipulating group composition might have unpredictable (and sometimes detrimental) effects on students' academic outcomes. In the context of the U.S. Air Force Academy, Carrell, Sacerdote, & West (2013) show that low-ability students placed into "optimally" designed peer groups perform significantly worse than comparable students who were randomly allocated to squadrons.[7] The explanation for this result is that the treatment changed the patterns of social interactions in ways that were key for student achievement. This evidence highlights how policy-induced patterns of social interactions may be a major obstacle to predicting the effects of altering peers' composition. Such concerns cast some doubt on the external validity of studies that randomly assign individuals to groups.

Third, this study provides the first successful example of how peer effects can be exploited in the design of public policies aimed at improving students' academic performance. In contrast to Carrell,

[3] An important exception is Fruehwirth (2013), who estimates spillover effects in academic outcomes in the context of a student accountability policy in North Carolina.

[4] Hoxby & Weingarth (2005) study the impact of peers' *lagged* achievement as opposed to *contemporaneous* achievement, which is the focus of this paper.

[5] See De Paula (2017) for a discussion of the different implications of endogenous and exogenous peer effects for the propagation of shocks into a network.

[6] See Sacerdote (2001), Cullen, Jacob, & Levitt (2006), Lyle (2007), Carrell, Fullerton, & West (2009), Duflo, Dupas, & Kremer (2011), and Carrell, Sacerdote, & West (2013) for examples of papers that use the random allocation of students to groups to estimate peer effects.

[7] In the U.S. Air Force Academy, incoming students are randomly allocated to squadrons. The design of "optimal" peer groups relied on estimating flexible reduced-form specifications of peer-effects in academic achievement using pre-treatment data. The objective was to maximize the outcomes of low-performing students.

Sacerdote, & West (2013), who focus on *exogenous* peer effects by randomly varying the *composition* of peer groups, we exploit the existence of *endogenous* effects within preexisting peer groups. Our results show that policies aimed at improving the bottom of the distribution have the potential to generate social-multiplier effects. Importantly, the findings indicate that it is possible to substantially improve academic outcomes *for all* with interventions targeted to the weakest. We believe that these considerations are important to inform any policy debate concerned with the allocation of public funds to education.[8]

Finally, by showing how the failure to consider general equilibrium effects might lead to an *underestimation* of the impacts of a policy, this paper also contributes to the policy evaluation literature. It is important to underscore that in our context, confining the consideration of the treatment effect to the *eligible* population would underestimate the benefits of the program by 47 percent. Thus, our findings underline the need to collect data on the entire local economy to fully appreciate policy effects. In addition, the results suggest the importance of experimentally manipulating individuals' treatment status within treatment units (schools in our setting) to identify social interactions.

The rest of this paper is organized as follows: Section 2 describes the remedying education intervention, the evaluation design, and the experimental results on the sample of low achievers. In Section 3, we discuss the issues related to the identification of peer effects and explain how we use the intervention to overcome these identification challenges. Section 4 presents the results. Section 5 addresses potential threats to identification, and discusses mechanisms and policy implications. Section 6 concludes.

2 The remedial literacy program

2.1 Setting

Public schools in Colombia, operate for a minimum of 165 days a year in either six or eight-hours shifts. One teacher typically teaches all subjects for a given grade. The primary school curriculum includes four main academic subjects: Spanish, mathematics, natural sciences, and social sciences. In addition to learning these subjects, students also study other subjects, including art, physical education, and technology. Although there are national guidelines regarding what children should achieve, schools and teachers are free of choosing pedagogical approaches and classroom strategies (Ministerio de Educacion

[8] At the macro level, achieving universal basic skills for *all* has the potential to generate increased and more equitable economic growth (Hanushek & Woessmann (2015)).

Nacional (2016)).

The remedial education program took place among third-grade students in public elementary schools in the municipality of Manizales in Colombia during three consecutive years (2015-2017). Manizales is a mid-size city in central Colombia. Approximately 13.8 percent of residents have incomes below the poverty line, and 6.9 percent of the municipality's residents live in rural areas. About 70 percent of the children in our sample can be considered socio-economically disadvantaged by SISBEN scores (the proxy mean tests used target social programs in the country).[9] In Manizales, about 78 percent of school-aged children attend public schools, and most children in our sample attended the school closest to their home. The municipality scored slightly above the national mean among third-graders in the 2016 national standardized language achievement tests (Pruebas Saber). However, almost 45 percent of students scored at or below the minimal-knowledge threshold in standardized official tests (Alcadía de Manizales (2017)). As a result, the local Secretary of Education, in partnership with a local NGO (Fundacion Luker) and the Inter-American Development Bank, implemented a remedial program to improve reading fluency among struggling third-grade students, and designed the evaluation of its effectiveness.[10]

2.2 Small-group tutorials for low-achieving students

The program provided students with 40-minute sessions three times a week for up to 16 weeks in the second half of the school year. The tutorials were conducted in small groups of up to six students and followed a simple structure. During each lesson tutors explained the objectives and activities, modeled the different exercises, and used both guided practice and student independent practice. The sessions used a curriculum designed and refined by international experts with support from a local team. The curriculum was based on a phonics approach. Lessons emphasized the ability to identify and manipulate units of oral language, the ability to recognize letter symbols and the sounds they represent, the ability to use combinations of letters that represent speech sounds, reading of words, and reading fluency of sentences and paragraphs. It also worked on vocabulary and strategies for reading comprehension.

The intervention targeted struggling readers who were identified using a measure of language devel-

[9] SISBEN scores are used to classify Colombian households in socio-economic strata. According to SISBEN scores, 30% of the students in our sample are in the lowest stratum (stratum 1), 40% in the second lowest stratum (stratum 2), and 30% are above these lowest two categories (strata 3,4,5 and 6), with most of them falling in stratum 3 (27%). This compares to an average of 34.4% in Colombia.

[10] Marinelli, Berlinski, & Busso (2021) provide further details regarding the intervention, the experiment, and its results on the target population.

opment. At the beginning of the school year, the Early Grade Reading Assessment (EGRA) was used to collect information on the following literacy subtasks: knowledge of letter sounds, reading of non-words, fluency of oral reading, and reading comprehension.[11] This information was used in each school to determine which students were eligible to participate in the tutorials. Eligibility was determined by local pedagogues based on the literacy skill level expected from a second-grade pupil and required that children could not correctly read more than 60 words out of the 132 words in a paragraph in the EGRA fluency of oral reading subtask. This sub-task relates to the "consolidated-alphabetic phase" in standard literacy acquisition models (Dubeck & Gove (2015)). Children in this phase are able learn new words through reading phonograms (or multi-letter patterns) rather than individual phonemes, and develop an increasing automatic sight word recognition which makes it easier for them to expand their vocabulary and reduces memory load (Ehri (1995)).[12]

Importantly for this paper, this strategy naturally generated two groups of students within the same class: a group of low-achieving students who were *eligible* to receive the intervention, and a group of higher-achieving students who were *not eligible*.

Throughout the paper, we define two students as peers if they are in the same class in the same school, rather than defining peers as those in the same grade, as many studies have done. We argue that the classroom-based peer definition offers a better approximation of how students interact in primary schools. For example, children in our sample spend at least 6 hours a day for roughly 165 days a year with their classmates, while the occasions to interact with other schoolmates are rather rare and mostly limited to playtime during recess. This is particularly important given that peer-effects estimates have been shown to greatly depend on the accuracy of the identification of relevant peers (Carrell, Fullerton, & West (2009)). For instance, Burke & Sass (2013) find evidence of peer effects at the classroom level but not at the grade-within-school level for elementary school children.

After collecting students' baseline data and determining which students were eligible to take part in the intervention, schools were randomly assigned to treatment in the following way: i) Schools were sorted based on how many low-achieving students were enrolled in third grade, and stratified in blocks of size two. ii) Within these strata one school was randomly selected to receive treatment, and the other one was selected to be a control. Low-achieving students in treatment schools participated in the remedial reading program, while those in the control schools carried on with their usual classroom-

[11] For more information on the EGRA test see Dubeck & Gove (2015).

[12] The eligibility criteria were slightly different in the first cohort. In this cohort, students were eligible if they scored in the bottom 25 percent of an equally weighted composite index of the following EGRA subtasks: reading of non-words, fluency of oral reading, and reading comprehension.

learning experiences. iii) Tutors were then randomly assigned to schools (one tutor per school), and, in schools with more than six eligible children, students were randomly assigned to equally sized tutorials.[13] This procedure was repeated each year of the intervention so that the *same* school could potentially be in a different treatment status from one year to the next.

Struggling readers in treatment schools were taken out of the classrooms during regular school hours. Tutorials took place in a designated school space at different times during the school day, and not specifically at the same time as literacy classes. The cap on tutorial size and the randomization of low-achieving students to different tutorials effectively meant that at any given point in time during the school day higher-achieving students would still be sharing the class with some of their lower-achieving classmates. This feature of the design is important because it allow us to rule out alternative explanations for the effects we find on higher achievers. We discuss this in greater detail in Section 5.

Tutors were hired each year of the intervention, and they were trained specifically to deliver the remedying program. The tutors were trained primary school teachers, psychologists, or audiologists with some teaching experience. There were no planned interactions between tutors and regular classroom teachers or non-eligible students, and qualitative interviews with principals confirmed that such interactions did not take place. Similarly, teacher interviews suggest that teachers in treatment schools were not more aware of the phonics approach compared to teachers in control schools. (We discuss teachers' practices and knowledge in more detail in Section 5.2). The sessions took place for 120 minutes each week. During that time the low-achieving students participated in the remedial intervention, and their higher-achieving peers continued receiving instruction using the standard curriculum. The higher-achieving students had no direct interaction with the remedial intervention program. Their only exposure was indirect, occurring through their interactions with low-achieving children with whom they shared the classroom every day and in all subjects.

Finally, it is important to stress that the experiment was designed to study the effect of the remedial tutorial sessions on the literacy skills of eligible children *only*, and not to uncover the potential spillover effects on their non-eligible peers. Nonetheless, the random assignment of the intervention across schools, and the fact that standardized test score data were collected for all students in the school allow us to study whether and how the academic outcomes of higher-achieving peers were affected by providing extra support to the weakest students in the class.[14]

[13] In the first year of the program, when there were more than six low-achieving students in the school, schools organized the compositions of the tutorials.

[14] While the intervention only targeted literacy skills, data were also collected for math achievement, as previous work demonstrates that literacy intervention could have positive impacts on other subjects by enhancing students' ability to

2.3 Data and descriptive statistics

The key outcome of interest is student academic achievement as measured by standardized language and math test scores on the Early Grade Reading Assessment and the Early Grade Math Assessment (EGRA and EGMA, RTI-International (2009)). Both tests were administered at the end of the school year by trained enumerators, who interviewed all students individually using a tablet. Our main outcome variable is the sum of correct answers across all reading and math subtasks standardized by the mean and standard deviation observed in the control group of each cohort.[15] We also report similarly defined literacy and math scores. In addition, we rely on information on child gender, age, and socio-economic status extracted from the administrative school records of the Integrated Enrollment System (Sistema Integrado de Matricula, SIMAT), the national database for the registration of students in public education in Colombia. This dataset also contains school-level information, which we used to compute class size.

[Figure 1 here]

Consistent with the evidence in Carrell & Hoekstra (2010) and Lavy, Paserman, & Schlosser (2011), in our sample we find that the test scores of higher-achieving students negatively correlate with the share of low achievers in their classroom. Figure 1 plots the relationship between end-of-the-year literacy and math scores of non-eligible students and share of low achievers prior to the intervention, in the sample of control schools only. Both for literacy and math, average achievement decreases monotonically with the share of low-achieving students.[16]

[Table 1 here]

Because the goal of this study is to understand the role that peers play in standard classroom settings (medium-sized, single-grade classrooms with one teacher), we restrict the analysis to randomization strata in which enrollment in third grade was larger than 20 students.[17] Table 1 presents summary statistics of the schools and children in our sample by treatment status. Panel A reports

follow instructional materials (Machin & McNally (2008); Machin, McNally, & Viarengo (2018)).

[15] Computing total score as the average between the math and reading scores does not change our main results.

[16] While there might be unobservable characteristics that simultaneously affect both the share of low achievers and the test scores of their higher-achieving peers, the figure controls for that type of selection by including school fixed effects. Importantly this relation is also robust to controlling for peers' SES status (captured by the social stratification classification scale used in Colombia to target social programs) as shown in Appendix Figure A1. This is suggestive that peers' achievement has an effect on individual academic performance over and above the effect of peers' background characteristics.

[17] The results are robust to using other thresholds. Appendix Figure A2 shows the robustness of the reduced form effects (as described below) on higher-achieving students to alternative thresholds.

school and class characteristics. Children's characteristics are reported separately for low-achieving students who were eligible to take part to the intervention, and higher-achieving students who were not eligible, in panels B and C respectively.

The treated and control groups are very similar in terms of observable characteristics, as one would expect given the randomized nature of the program. Low-achieving students and higher-achieving students are clearly different with respect to their school achievement levels. The scores of eligible children are significantly lower than those of non-eligible children. The magnitude of these differences is 47 points in literacy (out of 240, with p-value = 0.000) and 5 points in math (out of 52, with p-value = 0.000). For literacy, this knowledge gap is comparable to a full year of learning in the control group.[18] Eligible students are also more likely to belong to a lower socio-economic class as measured by SISBEN scores (the difference is 0.08, with a p-value = 0.000), and are on average slightly older than non-eligible students (the difference is 0.15, with a p-value=0.000).

2.4 Experimental Results

[Table 2 here]

Marinelli, Berlinski, & Busso (2021) present the evaluation of the program among the population of eligible students, showing that the overall literacy score of low-achieving students in treated schools improved compared to similar students in control schools. Table 2 replicates the main experimental results. At endline, the scores of low-achieving students in treatment schools were 0.362 of a standard deviation higher in literacy than those of similar, low-achieving students in control schools (column 1, Panel A of Table 2). The coefficient is virtually unchanged when including individual or class-level control variables (columns 2 and 3). Column 4 shows the results in the complete (unrestricted) sample. Panel B of Table 2 also reports a positive but not statistically significant result on math scores. The results in Table 2 clearly highlight that the intervention was effective in increasing the test scores of low achievers. This is important because it provides us with a source of exogenous variation in peers' contemporaneous test scores that we can exploit to study peer effects in academic achievement (see Section 3).

Marinelli, Berlinski, & Busso (2021) further show that the effect of the intervention is homogeneous in key respects. There are fairly constant quantile treatment effects. There seems to be no significant

[18] To compute this gap we use data collected at the beginning and at the end of the school year in control schools, and we take the difference in average test scores at these two different points in time.

heterogeneity among students who attended smaller or larger tutorials, or those who had comparatively worse or better tutorial peers, or those who were in more homogeneous or more heterogeneous tutorial groups, in terms of baseline reading ability. They also show an important dosage response, so that higher tutorial attendance predicts higher literacy scores.

3 Identification strategy and methodology

As discussed in Manski (1993), credibly identifying and quantifying peer effects pose important empirical challenges. In this section, we first describe these challenges, and then explain in detail our identification strategy.

First, the *simultaneity or reflection problem* arises as students affect each other, so that there is no exclusion restriction that can be used to distinguish the effect the individual has on the group from the effect the group has on the individual. Second, correlated unobservables plague identification when not all relevant group or individual characteristics are observed. These unobservables can generate a spurious correlation in outcomes that do not represent causal effects (Lyle (2007)).[19] Third, endogenous group membership is an issue because individuals self-select into peer groups or classrooms in a manner that is unobserved by the researcher. Positive selection frequently occurs with similar people joining the same group. This phenomenon, known as *homophily,* implies an upward bias in the estimated magnitude of peer effects.[20]

Previous research has tried to overcome these issues by including an extended set of controls for students and school characteristics. This often means using student and school fixed effects, or exploiting the naturally occurring variation in cohort composition over time within a school to deal with selection into peer groups.[21] Because results could still be biased, a second set of studies has exploited the random or quasi-random variation in peer-group composition to identify peer effects.[22]

[19]This would still be a problem in individuals were randomly assigned to groups. For example, in the educational context, this could be interpreted as a teacher fixed effect. Randomization of students into classes would still imply that students within the same class are exposed to the same teacher; a positive correlation in outcomes could be the results of same teacher exposure rather than a causal effect of peers.

[20]McPherson, Smith-Lovin, & Cook (2001) report the relevance of this phenomenon to explain the formation of social ties in a wide range of contexts, including marriage, work advice, information transfer, and friendship. In the educational context, Carrell, Sacerdote, & West (2013) report that students are more likely to interact with peers of similar ability and form homogeneous subgroups within the class, even when they are randomly assigned to classes.

[21]Studies that use this strategy include Hoxby (2000), Hanushek, Kain, Markman, & Rivkin (2003), Lefgren (2004), Hoxby & Weingarth (2005), Carrell & Hoekstra (2010), Lavy, Paserman, & Schlosser (2011), Burke & Sass (2013), and Card & Giuliano (2016).

[22]Papers that use the random assignment of students to groups include Sacerdote (2001), Cullen, Jacob, & Levitt (2006), Lyle (2007), Carrell, Fullerton, & West (2009), Duflo, Dupas, & Kremer (2011), and Carrell, Sacerdote, & West (2013). Other studies have used natural experiments as source of exogenous variations in peer composition. For examples, see Angrist & Lang (2004), Cipollone & Rosolia (2007) and Imberman et al. (2012).

While these papers credibly tackle the issue of self-selection, they effectively answer the question, "What would happen if individuals were randomly assigned to peer groups?". Whether the findings of these studies are generalizable to naturally occurring peer groups is not obvious. In particular, the patterns of social interactions that exist in these two different types of groups may differ fundamentally, resulting in different effects of peers on individual outcomes.

This is not mere theoretical speculation. The results in Carrell, Sacerdote, & West (2013) directly speak to this issue. Using the random allocation of cadets to squadrons in the U.S. Air Force Academy, Carrell, Sacerdote, & West (2013) estimate flexible reduced-form specifications of peer effects in academic achievement. Using these estimates, they allocate incoming students to squadrons to maximize the achievement of lowest-performing students. Surprisingly, their findings show that low-ability students placed into these "optimally" designed peer groups performed significantly worse than comparable low-ability students who were randomly allocated to squadrons. The explanation for this puzzling result is that the treatment changed the endogenous patterns of social interactions in ways that were key for student achievement. The authors show that within their optimally designed groups, low-performing students avoided interacting with high achievers (the very students they intended them to interact with), and instead formed more homogeneous subgroups. This evidence highlights how policy-induced patterns of social interactions may be a major obstacle to predicting the effects of altering peer groups. The findings cast some doubt on the external validity of studies that randomly assign individuals to groups.

Finally, a small but growing literature exploits *partial population experiments* (Moffitt (2001)), to study peer effects in naturally occurring groups. This approach uses the experimentally induced variation in the outcomes of a subset of individuals in the relevant peer group to identify peer effects for the non-treated individuals. This approach has been used to study labor market outcomes (Hesselius, Nilsson, & Johansson (2009)), financial decisions (Bursztyn, Ederer, Ferman, & Yuchtman (2014)), retirement plan decisions (Duflo & Saez (2003)), social program participation (Dahl, Løken, & Mogstad (2014)), and healthy behavior (Centola (2010)). Only a very few papers have used this approach in the context of education, and most of them have looked at peer effects in school enrollment rather than academic achievement (Bobonis & Finan (2009), Lalive & Cattaneo (2009) and Angelucci, De Giorgi, Rangel, & Rasul (2010)).[23] Our approach is similar to these studies in that we exploit a randomized

[23]So far as we are aware, the only other paper that uses a partial population experiment to look at peer effects in achievement is Boozer & Cacciola (2001) in the context of the Tennessee Student-Teacher Achievement Ratio experiment (Project STAR). However, that study analyzes peer effects in groups that are randomly assigned. Therefore, the concerns of external validity raised above are still valid for that study. Using a design similar to the one used in this paper, Johnson

control trial designed to improve reading fluency among low-performing students to study academic achievement of their *non-treated*, higher-achieving peers. The remedying education program exploited in this paper offers a unique opportunity to analyze whether an exogenous increase in the test scores of peers within a class increases individual achievement.

The essence of our identification strategy can be more easily understood by considering the following system of equations. For simplicity, imagine that the reference group (i.e., the class) is only composed of three students: A, B and C, where A and B are higher achievers, and C is a low-achieving student. Then we can write:

$$\begin{aligned} y_{A,G} &= \alpha + \rho\left(\frac{y_B + y_C}{2}\right) + \delta x_A + \gamma\left(\frac{x_B + x_C}{2}\right) + \theta\omega_G + \epsilon_{A,G} \\ y_{B,G} &= \alpha + \rho\left(\frac{y_A + y_C}{2}\right) + \delta x_B + \gamma\left(\frac{x_A + x_C}{2}\right) + \theta\omega_G + \epsilon_{B,G} \\ y_{C,G} &= \alpha + \rho\left(\frac{y_A + y_B}{2}\right) + \delta x_C + \gamma\left(\frac{x_A + x_B}{2}\right) + \theta\omega_G + \tau T_G + \epsilon_{C,G} \end{aligned} \tag{1}$$

where $y_{i,G}$ is the academic achievement of student i in group G, x_i are individual observable characteristics, ω_G are observable group specific characteristics, and $\epsilon_{i,G}$ is an error term. Notice that treatment T_G varies randomly across groups, but there is no change for any higher-performing student. In the terminology of Manski (1993), ρ is the endogenous effect emanating from peers' contemporaneous outcomes, while γ is the exogenous effect from peers' background characteristics. The focus of this paper is the endogenous peer effect.

The random assignment of the treatment overcomes the identification challenges in the following ways. First, it solves the reflection problem because the experiment induces exogenous variation in the outcomes of the low-performing child (student C) without *directly* affecting higher-achieving students (A and B). Second, randomization implies that the treatment is orthogonal to all observable and unobservable characteristics (x_i, ω_G, and $\epsilon_{i,G}$), solving the problem of correlated unobservables.[24] Finally, because peer groups are established before the policy change and fixed throughout the experiment, endogenous group membership is not an issue. We can think of peer effects as being conditional on any selection into groups that might have taken place prior to the experiment. For this reason, the identification strategy allows us to identify the effect of peers in naturally occurring groups.

et al. (2019) study the effects of small group tuition for 5-year-old pupils in English schools and find spillover effects to control students in treatment schools. Johnson et al. (2019) do not use the experimental variation to estimate peer effects in academic achievement.

[24] The evidence in Table 1 shows balance between the treatment and control groups in terms of observable characteristics.

We can identify the causal effect of the program τ by regressing $y_{C,G}$ on T_G. The endogenous peer effect ρ is identified by regressing $y_{i,G}$ (for $i = A, B$) on T_G and scaling by $\hat{\tau}$. This is equivalent to an instrumental variable strategy that uses T_G as instrument for average peer achievement in the equation of higher-achieving students.

Formally, we estimate the following linear-in-means model of peer effects using two-stage least squares (2SLS) on the sample of *higher achievers* only (i.e., on the sample of students who were *not eligible* for the remedying intervention):

$$y_{icst} = \rho \bar{y}_{-icst} + X_{icst}\beta + \omega_s + \lambda_t + \epsilon_{icst} \tag{2}$$

where i is the student, c is the class, s is the school, and t is the cohort. The outcome variable y_{icst} is the test score of a student (expressed in standard deviations of the distribution of scores in control schools), $\bar{y}_{-icst}$ is the average contemporaneous score of her peers, X_{icst} is a vector of child/class-specific characteristics and ω_s, λ_t are school and year fixed effects, respectively. The repeated randomization of schools into treatment and control groups over time allows for the inclusion of school fixed effects in equation (2). This allows us to control for time-invariant determinants of student achievement at the school level, effectively controlling for the possible selection of students into schools.[25] We instrument $\bar{y}_{-icst}$ in (2) using the school treatment status.[26] Given the potential for error correlation across students within a given peer group, we cluster all standard errors at the class level. The coefficient ρ is the *endogenous* peer effect (Manski (1993)). This captures the effect of peers' contemporaneous test scores on individual achievement.[27]

Our identification strategy rests on the assumption that the treatment did not have any *direct* impact on the *non-treated*. This effectively means we have one variable that can be excluded from (2) while generating random variation in peers' average score. One potential concern is that by physically removing low-performing children from the classroom, higher-performing students may have experienced a reduction in class size that *directly* affected their test scores. In Section 5, we discuss

[25] In practice, it could also be that students are not randomly assigned to classes within a school. As discussed earlier, this does not pose a challenge to our identification strategy, and the effects we estimate can be thought as being conditional on any selection into groups that might have taken place prior to the experiment. Empirically there is little evidence to support the hypothesis that tracking within schools was relevant in our context. For example, as we would expect in the absence of tracking, the within-school variation accounts for 17 to 34 percent of the total variation in the share of low-achieving students in any given year.

[26] Using the class treatment status does not make any difference because randomization took place at the school level. Therefore all classes within the same school experienced the same treatment.

[27] The existence of endogenous peer effects in the production function for test scores can be micro-founded using an effort game in the classroom, in which students' effort is determined jointly with peers' effort (see Fruehwirth (2013) and Tincani (2017)).

this and other potential threats to identification, and we perform several robustness checks to address these potential concerns.

4 Results

As shown in Section 2.4, the intervention generated experimentally induced variation in the outcomes of a subset of the students within the class. This result is the basis of our 2SLS strategy. We start by presenting graphical and regression-based evidence of the reduced-form effect of being in a treatment class on the sample of higher-achieving students in Section 4.1. Then, in Section 4.2 we estimate linear-in-means models of peer effects; we regress non-eligible students' test scores on the average *contemporaneous* score of their peers. We also explore whether these effects are heterogeneous depending on a student's baseline achievement.

4.1 Reduced-form evidence

To assess the *indirect* effect of the remedial education program on higher-achieving students, in Figure 2 we plot, separately for treated and control schools, students' end-line test scores as a function of baseline scores using a second-order polynomial. In each graph, we plot local averages and polynomial fits estimated separately for the treatment and control groups. For comparison purposes, we start by presenting the test scores of low-achieving students, those that were directly targeted by the intervention (see Panel A). Perhaps unsurprisingly given the results in Table 2, the fitted values in treatment schools are consistently above those in control schools for this sample of low achievers.

[Figure 2 here]

Panel B illustrates the "reduced-form" effect of being in a treatment classroom for those students who were not eligible to receive the intervention because their baseline test score fell above the eligibility cutoff. Surprisingly, the same general picture observed for eligible children emerges for the non-eligibles students, too. Higher-achieving students in treatment schools systematically outperform similar students in control schools. This is true for all quantiles of the baseline achievement distribution, even if there is some suggestive evidence that the effects are stronger at higher quantiles. (We return to this point in Section 4.2.) As we would expect, the magnitude of the difference in test scores between students in treatment and control schools is smaller for higher-achieving students than for low-achieving students.

[Table 3 here]

In Table 3 we present the reduced-form estimate of the impact of the intervention on higher-achieving students. In column 1 we regress non-eligible students' outcomes on an indicator variable that takes the value of one if her school was in the treatment group, and zero otherwise. In columns 2 and 3 we include additional individual and class controls. Panel A reports the results for literacy, while panels B and C report those for math and total scores, respectively.

Consistent with the results shown in Figure 2, the literacy scores of high-achieving students in the treated schools were 0.108 of a standard deviations greater than the scores of similar students in the control schools. (The p-value of the difference is 0.064.) The effect is slightly larger when we control for individual characteristics (0.112 of a standard deviation, with a p-value of 0.053) and class characteristics (0.118 of a standard deviation, with p-value of 0.04). Similarly, we find that the total test scores of students whose peers' were treated increase by 0.112 of a standard deviation compared to students in the control group. (The p-value of the difference is 0.046.) This effect increases to 0.120 of a standard deviation when we include individual and class characteristics (p-value of 0.030). These effects are sizeable and represents roughly 30 percent of the treatment effect on eligible children. This result is economically meaningful, and its magnitude can be compared to a more commonly proposed school-level reform: tracking by prior achievement. Duflo, Dupas, & Kremer (2011) find that tracking raises literacy scores and total scores by 0.198 and 0.139 of a standard deviation, respectively, for students in both upper and lower tracks.

We find small and non statistically significant effects in math (similar to the results shown in Table 2). This is reassuring and gives us confidence that these effects are indeed driven by peer-to-peer learning. We discuss this issue in more detail in Section 5. Given that we do not find any reduced-form effect for math scores for non-eligible children, and that the results for treated students are negligible (as shown in Table 2), we focus on literacy and total scores.[28]

4.2 Peer effects in academic achievement

We now turn to the estimation of linear-in-means models of peer effects by estimating equation (2) on the sample of higher-achieving students only. Table 4 reports the OLS and 2SLS estimates. The OLS results in Panel A show that a one-standard-deviation increase in peers' contemporaneous achievement

[28] While the point estimates for maths are substantially smaller than those for literacy (by about one third) for both eligible and non-eligible students, they do not necessarily imply a null effect and the lack of a significant effect might be due to a lack statistical power to detect small effects on this outcome.

is correlated with an increase in literacy scores by 0.535 of a standard deviation (column 1). The result for total scores, shown in column 4, implies that a one-standard-deviation increase in average peers' scores is associated with an increase in individual achievement by 0.56 of a standard deviation.

[Table 4 here]

In panels B and C we report the first and second stage for the 2SLS model that uses the treatment status as an instrument for peers' average contemporaneous scores (the reduced-form was reported in Table 3, and is therefore omitted here). We have a very strong first stage: average peers' literacy score is 0.159 of a standard deviation higher in treatment classes compared to control classes (p-value = 0.007). By dividing the reduced-form coefficient (column 2 of panel A in Table 3) by the first-stage coefficient, we obtain an estimate of the peer effect parameter in equation (2). The 2SLS coefficient in column 1 implies that a one-standard-deviation increase in peers' contemporaneous achievement increases own achievement by 0.679 of a standard deviation. Column 4 reports the results for total scores, which are very similar, and imply that a one-standard-deviation increase in average peer end-line test scores would increase the test score of a student by 0.704 of a standard deviation.[29] These effects are comparable to those found in previous work. For instance, Boozer & Cacciola (2001) estimate an effect of 0.92 of a standard deviation for third-grade students, while Lavy & Schlosser (2011) find a peer coefficient of 0.84. Using data from the Project STAR experiment, Whitmore (2005) finds that peers' test scores increase the individual score by 0.6 of a standard deviation.[30]

Because the previous literature has found evidence of non-linearities in peer effects (see Sacerdote (2001), Burke & Sass (2013), Tincani (2017)), we investigate whether the same is true for endogenous peer effects. To examine this issue, we split the sample of non-eligible children using three terciles of the baseline achievement distribution, and estimate separate models for these three sub-samples. Table 5 reports the first-stage and second-stage regressions separately for students in the first, second, and third terciles of the baseline distribution of the outcome variable. For comparability purposes, in Panel A we report the second-stage coefficients from Table 4.

The results are consistent with the notion that students at the top of the achievement distribution benefit the most from improvements in their peers' outcomes.[31] We find that the peer-effect coefficient

[29] As pointed out by Duflo, Dupas, & Kremer (2011), these results come from variation in peers' average achievement that are smaller than one standard deviation, so the extrapolation to one standard deviation might not be precise if the effects are non-linear.

[30] De Giorgi, Pellizzari, & Redaelli (2010) also estimate endogenous peer effects in the context of choice of a subject major in university; it is worth noting that both we and they find that the 2SLS coefficient is larger than the OLS coefficient.

[31] By contrast, some papers that look at heterogeneous peer effects in academic achievement using the random allocation

monotonically increases with a student's baseline achievement quantile. Students just above the eligibility cutoff seem to be less affected by their peers compared to students at the top of the distribution. For these students a one-standard-deviation increase in peers' contemporaneous score increases own literacy scores by 0.777 of a standard deviation, and total scores by 0.832 of a standard deviation (Panel C of Table 5).[32]

[Table 5 here]

While we cannot estimate the effect on the lowest-achieving students – because these students were *directly* affected by the program – we find evidence that the endogenous peer effect is stronger for highest-performing students compared to "average"-performing students.[33]

5 Discussion

5.1 Threats to identification

As discussed in Section 3, our identification strategy rests on the assumption that the intervention does not directly effect learning outcomes of higher-achieving students in treatment schools. If this were not the case there would not be any source of exogenous variation in average peers' scores that we could use to implement an instrumental variable strategy. While this assumption is not testable, in this section we do our best to rule out possible alternative mechanisms that could explain the increase in test scores that we observe for higher-achieving students.

5.1.1 Class size

One potential concern is that by physically removing low-performing children from the classroom, higher-performing students experienced a reduction in class size which had a *direct*, positive impact on their test scores. We provide several pieces of evidence against this interpretation.

of students to peer groups find that high-performing students are less affected by peers' scores than lower-performing students. For example, Carrell, Fullerton, & West (2009) find that the peer-effect coefficient is larger for students in the bottom third of the academic ability distribution (even though they cannot reject the equality of the coefficients). A similar result is reported in Booij, Leuven, & Oosterbeek (2017). Is it important to note that these papers only estimate a composite parameter that incorporates both the endogenous and exogenous peer effects.

[32] The fact that that the standard errors are similar across subsamples provides evidence that the insignificant effects on the lower quantiles stems from the low magnitude of the estimates, not from a lack of statistical power.

[33] We have also tried expanding equation (2) to allow the peer coefficient to vary with students' baseline achievement levels, by estimating: $y_{icst} = \rho_1 \bar{y}_{-icst} \times Q_1 + \rho_2 \bar{y}_{-icst} \times Q_2 + \rho_3 \bar{y}_{-icst} \times Q_3 + X_{icst}\beta + \omega_s + \lambda_t + \epsilon_{icst}$ Where Q_1, Q_2 and Q_3 are indicator variables taking the value one if child i falls in the first, second, or third tercile of the baseline achievement distribution. The results, shown in Appendix Table A1, follow the same patterns as those shown in Table 5, but they are somewhat larger for children in the third tercile of the baseline distribution. In that model, we test and reject the hypothesis that $\rho_1 = \rho_2 = \rho_3$.

First, higher-performing students experienced only a modest reduction in class size, and for only a minimal amount of instruction time. This is because there was one single tutor per school and tutorial size was capped at six. Moreover low-achieving students from the same class were randomized into different tutorial groups that took place at different times during the school day.

[Figure 3 here]

Figure 3 shows the distribution of same-class students assigned to the same tutorial and the implied reduction in class size. In over 20 percent of classes only one student was assigned to the *same* tutoring group, and in more than 75 percent of classes fewer than four low-achieving students were randomized into the same group (Panel A).[34] The implied class size experienced by regular students, shown in Panel B, was composed of roughly three fewer students on average (out of an average class size of 31 students). Moreover, this reduction took place for only 40 minutes a day, three days a week, for a period of 16 weeks compared to the whole academic year (as in most studies on class size). This means that the class-size reduction experienced by regular students lasted for roughly 32 hours out of almost 1,000 yearly school hours.

The paper that documents the largest class size-effects in the literature is Urquiola (2006). The paper finds that reducing class size by on average nine students increases test scores by between 0.16 and 0.3 of a standard deviation. (Some of the results derive from up to three years of smaller class sizes).[35] In comparison, the reduction in class size in our experiment is substantially smaller, and lasted for a significantly shorter period of time. A back-of-the-envelope calculation implies that our reduction in class size predicts at most an increase in the test scores of regular students in the range of 0.005 to 0.01 of a standard deviation. Thus, any potential effect from class-size reduction would explain at most a tenth of the reduced-form effect found in Table 3; therefore class size is unlikely to be the driving force underpinning our results.

Second, an additional piece of evidence against the class size story comes from the lack of reduced-form effect on math test scores. The remedying tutorials did not take place specifically during regular literacy hours. For this reason, if the reduction in class size were to be the main driver behind the reduced-form effect of the intervention on higher-achieving students' test scores, we would expect to

[34]Notice that while the size of the tutorial groups was capped at six, our records show that there is one school where the actual size was of the tutorial was increased to seven.

[35]The results reported in Urquiola (2006) do not come from a randomized controlled trial (RCT). The only study we are aware of that uses an RCT to looks at the effect of reducing class size on learning outcomes in a developing country was conducted by Duflo, Dupas, & Kremer (2015), who find no significant test gains for students exposed to a smaller class size.

see an impact on this outcome as well. The fact that we did not find any economically meaningful and statistically significant effect on math scores for regular students in Table 3 (point estimate of 0.034 with an associated standard error of 0.049) rules out large effects coming from a reduction in class size.

Third, we estimated reduced-form impacts on higher achievers, separately for classes that experienced larger or smaller reductions in size (by virtue of the random assignment of eligible children to tutorial groups). The results (reported in Appendix Table A2) show that the effects are homogeneous along this margin, providing additional support against this interpretation.

5.2 Mechanisms

The term peer effect is generally used as an umbrella term that comprises *any externality*, implying that peers' outcomes have an impact on an individual's outcome. Both *direct* and *indirect* effects are peer effects. This effectively includes: i) peer-to-peer learning, ii) student misbehavior, and iii) teacher practices (Sacerdote (2011)). With the exception of Lavy, Paserman, & Schlosser (2011), we are not aware of other studies that have attempted to empirically separate these channels. In this paper we explore these alternative mechanisms because such distinctions might be key for the design of optimal education policies. It is important to stress that that these alternative explanations only affect the interpretation underlying the existence of peer effects, but do not invalidate our identification strategy.

5.2.1 Teacher responses and student misbehavior

While we are confident that our identification strategy does not suffer from any of the identification issues described in Section 3, it does not allow us to disentangle the effects coming from student-to-student interactions from those that stem from teachers' behavior. In particular, it might be that teachers changed their practices in ways that are key for student achievement.

To gauge the importance of teacher's behavior we use two alternative strategies. First, we note that within the same school different classes have different *shares* of low -achieving students, and over time the share of low-achieving students in a school varies.[36] Therefore, in the group of treatment schools, we have variation in the class share of treated students. We exploit this source of variation to implement an instrumental variable strategy in *treatment schools only*. This strategy is similar to the one previously described, but instead the average score of peers ($\bar{y}_{-icst}$ in equation (2)) is

[36] As we would expect in the absence of tracking, most of the within-school variation in our data comes from variation over time, rather than variation between classes in the same time period. In any given year, the between-school variation accounts for 66 to 83 percent of the total variation in the share of low-achieving students.

instrumented using the share of eligible (hence, treated) students. Identification here is achieved using (i) idiosyncratic variation in the proportion of low achievers within a school over time, and (ii) between-class variation in the proportion of low-achieving students within the same school. We further control for average achievement at baseline in the class, so that we effectively compare classes that are similar in terms of average baseline performance.

By considering *treatment schools only*, we ensure that all teachers are being exposed to the same "treatment" (the remedying intervention), so that the effects on non-eligible students cannot be explained by teaching practices that change *because* of the treatment.[37] This approach operates under the assumption that idiosyncratic variations in the share of low-achieving students (controlling for average baseline achievement and school fixed effect) within a school and over time do not systematically affect teacher practices in ways that matter for higher-achieving students' test scores; thus, this identification strategy allows us to tease out the peer effect coming exclusively from student-to-student interactions, net of any teacher response.

[Table 6 here]

The results are presented in Table 6. Panel A shows the OLS results, while panels B and C report the reduced-form and first-stage results. Conditioning on school fixed effects and average class achievement at baseline, increasing the share of treated students by 10 percentage points increases the average peer score by 0.077 of a standard deviation (p-value = 0.003). By dividing the reduced form by the first-stage coefficient, we calculate a peer-effect coefficient of 1.025 (Panel D of Table 6). This is not statistically different from the value of 0.705 found in Table 4. As a falsification test, we also estimated the same model in the sample of control schools. In the control group, we would expect a small and non-significant first stage because no remedial education intervention took place in these schools. We find this to be the case. Conditional on average baseline achievement in the class, a 10 percentage point increase in the share of low-achieving students translates in a non-significant change of 0.007 of a standard deviation in the average end-line score of non-eligible students' peers.[38]

[37] A more subtle issue is that teachers' practices might still be affected by the share of low-achieving students in the class. While we do not think this is a very compelling story, the following example illustrates a scenario in which our strategy would not effectively control for teacher responses. If, controlling for school fixed effects and the average achievement of the class, a teacher were to change her behavior when confronting a class that included a 10 percent share of low achievers as opposed to one including 20 percent of low achievers, then we would not be able to separately identify the effects of peers from those of teacher practices. Notice that by including school fixed effects we effectively control for differences in teaching strategies between schools. In our regressions we also control for mean achievement, so that any teacher response operating through that margin (as proposed in the model by Duflo, Dupas, & Kremer (2011)) would be taken into account. Thus, our approach would only miss potential impacts of this channel if teachers were acting on the share of low achievers, rather than the mean level of student achievement.

[38] The results are available from the authors upon request.

This strategy allows us to rule out some particular types of teacher responses that could explain our findings. Notably, any change in behavior that occurs because of the treatment would be taken care of by this identification strategy. This result is consistent with the peer effects we identify being driven by *direct* peer-to-peer interactions rather than by changes in teachers' educational practices in our setting.

To further rule out effects stemming from teachers' responses, we present direct evidence from a teacher survey that was administrated in a subsample of schools in our study sample. The survey included a set of items adapted from the teacher section of the Patterns of Adaptive Learning Scales (PALS). These scales are used to evaluate the teachers' outlooks on the school goals, approaches to teaching, and teaching efficacy (Midgley et al. (2000)). We focus on three subscales of the PALS: (i) "Performance approaches" refer to the strategies used by teachers to convey to students that the purpose of engaging in academic work is to demonstrate competence. (ii) "Teacher efficacy" relates to teachers' beliefs that they are contributing significantly to the academic progress of their students, and that they can effectively teach to all students in their class. (iii) "Student bad behavior" captures the extent to which teachers have to deal with student misconduct during school hours. Using the items from each of these subscales, we construct a composite using principal component analysis, which we then standardize to have a mean of zero and standard deviation of one in the control group.

We also investigate whether regular teachers in treatment schools did learn about phonic practices from the hired tutors. To this aim we use survey questions inquiring directly about the phonics approach used in the tutorials, e.g. "Phonological awareness is indispensable for reading". Using the answers to these questions, we construct a composite using principal component analysis.[39]

To analyze whether there are differences in the behavior and knowledge of the teachers in treatment and control schools, we regress each outcome on a treatment-indicator variable, controlling for teacher characteristics.[40] Table 7 shows the results.

[Table 7 here]

For all three PALS outcomes we cannot reject the null hypothesis of equality between teachers in treatment and control schools. However, the sample size might be too small to detect statistically significant differences. We find that the point estimates in columns 1 and 2 are very small in magnitude,

[39]Unfortunately the sample of teachers that answered this part of the teacher survey further reduces. Importantly, response rates are not related to treatment (a regression of response rate on the treatment dummy has a small, negative and not statistically significant coefficient (the difference is -0.01, with a p-value of 0.885)).

[40]The results when we do not include teacher characteristics are virtually identical and are not reported.

while the point estimate from column 3 is substantially larger, suggesting an effect of over 0.3 of a standard deviation on students' misbehavior. This provides some suggestive evidence that teachers reported having to deal with students' misbehavior more often in control schools than in treatment schools. This result is consistent with previous work showing that classroom disruptions decrease with an increase in students' ability (see Carrell & Hoekstra (2010) and Lavy, Paserman, & Schlosser (2011)). Similarly, we do not find evidence that teachers in treatment schools have any superior knowledge about the phonics approach compared to teachers in control schools.

The lack of evidence of teacher responses is consistent with previous work demonstrating the difficulties of affecting teacher's behavior, even with interventions designed to do just that. A rigorous evaluation of a incentives program in Kenya - that directly targeted teachers - finds that "there is little evidence that teachers in the program schools increased efforts to reduce dropouts or promoted broad acquisition of human capital" (Glewwe, Ilias, & Kremer (2010)). Similarly, evaluations of teacher development programs fail to find robust effects on teachers's behavior (Loyalka, Popova, Li, & Shi (2019)).[41] One important exception is Duflo, Dupas, & Kremer (2011). In the context of tracking, the authors find that teachers in tracking schools increase their effort.

In thinking about these results, it is important to keep in mind that the intervention considered in this paper did not involve teachers in any way. No direct contact occurred between teachers and tutors, and teachers in treatment schools were not more likely to know about the phonics approach that tutors used during the tutorials. Moreover, there was no change in the composition of the student body; throughout the duration of the experiment, teachers kept teaching to the exact, same group of students. We therefore do not believe that teachers effort or pedagogical instructions are a major driver of the results find for non-eligible students, and find little very empirical support for this explanation.

5.2.2 Peer-to-peer interactions

The previous section provides suggestive evidence that the peer effects we estimate do not stem from a change in teachers' effort, and are thus consistent with *direct* peer-to-peer interactions. We now provide further evidence in favor of this interpretation. To this aim, we investigate whether higher-achieving students in treatment schools are affected by *improvements* in their lower achieving peers. The idea is that if peer-to-peer learning is important in explaining the effects we estimate, then we

[41] Popova, Evans, Breeding, & Arancibia (2021) argue that teacher development programs *can* be effective if they have three specific characteristics: (i) linking participation in professional development to promotion or salary increases, (ii) having a specific subject focus, and (iii) allowing teachers to enact lessons as part of the training.

should observe larger improvements for non-eligible students in classes where eligible students improved more, compared to similar students in classes where eligible students improved less. We estimate the following specification in treatment schools only:

$$y_{icst} = \gamma \Delta_c^E + X_{icst}\beta + \omega_s + \lambda_t + u_{icst} \tag{3}$$

where y_{icst} in the outcome of (non-eligible) student i, in class c of school s, in cohort t, Δ_c^E is the average change in test scores of low-achieving students in class c, and all other variables have been previously defined. Because Δ_c^E might not be exogenous in (3), we instrument it with average tutorial attendance among eligible peers. Marinelli, Berlinski, & Busso (2021) find that dosage played an important role in explaining gains in the population of eligible students. Moreover, variation in tutorial attendance varied because of reasons unrelated to the performance of non-eligible students, and had to do with specific program implementation features, such as the availability of make-up sessions.

[Table 8 here]

The results from this specification are shown in Table 8. The OLS results in Panel A suggest a positive correlation between changes in the scores of lower-achieving peers and an individual score, and Panel C report a strong first stage, so that improvements were larger when attendance was higher. The 2SLS results in Panel D are consistent with direct peer-to-peer interactions. Specifically, the coefficients suggest that within the sample of treatment schools only, the test scores of non-eligible students were higher in classes where eligible students improved more (the p-values are 0.031 and 0.049 for literacy and total scores).

5.3 Implications

The findings in this paper imply that failing to consider the indirect effects of the remedying intervention on non-eligible students underestimates the true treatment effect for the overall student population. Consider the following back-of-the-envelope calculation: The intervention cost USD 89 per eligible student in 2016 (Marinelli, Berlinski, & Busso (2021)).[42] Using our results from Table 2, we calculate that for every USD 100 spent, low-achieving students' test scores increased by 0.406 of a standard deviation, *and* higher achievers' test scores also increased by 0.121 of a standard deviation.[43]

[42] The authors use the Ingredients Approach to compute these costs (Dhaliwal, Duflo, Glennerster, & Tulloch (2013)).

[43] For eligible students, this is given by $0.362 \times (\frac{100}{89}) = 0.407$. For non-eligible students, using our results from Table 3 this is given by $0.108 \times (\frac{100}{89}) = 0.121$.

Given that there are three times more non-eligible students than eligible students, this translates into an additional increase of 0.363 of a standard deviation in test scores for every USD 100 spent. Therefore, by our calculations, a failure to consider the impact on the program on higher-achieving students in treated schools underestimates the effect of the remedying education intervention by 47 percent.[44]

The results in this paper thus underline the need to collect data on the entire local economy to fully appreciate policy effects and to correctly compute the *returns* to remedial education policies. Endogenous peer effects lead to a social-multiplier effect that amplifies the total output of a program. From a methodological point of view, our findings emphasize the importance of experimentally manipulating individuals' treatment status within treatment units (in our setting, schools) to identify social interactions.

6 Conclusions

In this paper, we analyze whether providing academic support to the lowest-performing students in a class affected their higher-performing peers. We examine the impacts from a randomized experiment that provided tutoring to students who had the lowest reading skills in Colombian schools. In treatment schools, students with low baseline reading scores were assigned to small, group tutoring classes during which they worked with a qualified tutor following a structured pedagogical curriculum. The randomization strategy naturally generates two groups of students within the same class: a group of low-achieving students who were *eligible* to receive the intervention, and a group of higher-achieving students who were *not eligible*. We can therefore study whether an exogenous change in the test scores at the bottom of the class translates into gains at the other levels.

The intervention was very effective in improving literacy skills in the sample of low-achieving students: average test score increased by 0.362 of a standard deviation by the end of the intervention. We find that the intervention improved the learning of everyone else in the class – regardless of previous literacy achievement levels. We compare the test scores of higher-achieving students after one academic year, finding substantially greater achievement across the board in treated schools compared in control schools. That is, in the treatment schools, higher-achieving students who did participate to the tutoring activies outperformed similar students in control schools by 0.108 of a standard deviation. This coefficient is sizable and represents roughly 30 percent of the treatment effect on the *eligible* students.

[44] This is given by $\frac{0.363}{(0.363+0.407)} = 0.47$.

Using the treatment-induced variation in peers' scores as an instrument for peers' outcomes, we estimate a linear-in-means model of peer effects, focusing particularly on their endogenous component. The random allocation of the treatment allows us to overcome the identification challenges that have plagued much of the previous literature on peer effects – namely selection, reflection, and correlated unobservables (Manski (1993)). We find strong evidence of peer effects in academic outcomes. Our results imply that a one-standard-deviation increase in peers' *contemporaneous* test scores increases individual reading scores by 0.679 of a standard deviation. We find evidence of non-linearities, with largest effects at the top of the ability distribution. We further rule out alternative mechanisms coming from a reduction in class size. We do not find evidence that teachers changed their effort or teaching practices. We find some suggestive evidence that some of the effect might be due to a reduction in students' misbehavior. Finally, we show that the effects are stronger in classes where eligible peers improved the most, consistent with *direct* peer-to-peer learning interactions.

This study provides the first successful example of how peer effects can be exploited in the design of public policies aimed at improving students' academic performance. Taken together, our findings suggest that policies aimed at improving the bottom of the achievement distribution have the potential to generate social-multiplier effects across the board. This indicates that it is possible to substantially improve the quality of education *for all* with relatively cheap and easy-to-scale interventions. The findings provide a strong rationale that underscores why society should care about improving the educational outcomes of the weakest. Moreover, at the macro level, achieving universal basic skills for *all* has the potential to generate increased and more equitable economic growth (Hanushek & Woessmann (2015)). These considerations are important to inform any policy debate concerned with the allocation of public funds to education.

References

Alcadía de Manizales. (2017). Boletín estadístico del sector educativo año 2017, Manizales, Caldas. *mimeo*.

Angelucci, M., De Giorgi, G., Rangel, M. A., & Rasul, I. (2010). Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of public Economics*, *94*(3-4), 197–221.

Angrist, J. D., & Lang, K. (2004). Does school integration generate peer effects? evidence from boston's metco program. *American Economic Review*, *94*(5), 1613–1634.

Bobonis, G. J., & Finan, F. (2009). Neighborhood peer effects in secondary school enrollment decisions. *The Review of Economics and Statistics*, *91*(4), 695–716.

Booij, A. S., Leuven, E., & Oosterbeek, H. (2017). Ability peer effects in university: Evidence from a randomized experiment. *The review of economic studies*, *84*(2), 547–578.

Boozer, M., & Cacciola, S. E. (2001). Inside the'black box'of project star: Estimation of peer effects using experimental data. *mimeo*.

Burke, M. A., & Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics*, *31*(1), 51–82.

Bursztyn, L., Ederer, F., Ferman, B., & Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, *82*(4), 1273–1301.

Card, D., & Giuliano, L. (2016). Can tracking raise the test scores of high-ability minority students? *American Economic Review*, *106*(10), 2783–2816.

Carrell, S., Fullerton, R. L., & West, J. E. (2009). Does your cohort matter? measuring peer effects in college achievement. *Journal of Labor Economics*, *27*(3), 439–464.

Carrell, S., Hoekstra, M., & Kuka, E. (2018). The long-run effects of disruptive peers. *American Economic Review*, *108*(11), 3377–3415.

Carrell, S., & Hoekstra, M. L. (2010). Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics*, *2*(1), 211–28.

Carrell, S., Sacerdote, B., & West, J. (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, *81*(3), 855–882.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194–1197.

Cipollone, P., & Rosolia, A. (2007). Social interactions in high school: Lessons from an earthquake. *American Economic Review*, *97*(3), 948–965.

Coleman, J. S. (1968). Equality of educational opportunity. *Integrated Education*, *6*(5), 19–28.

Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica*, *74*(5), 1191–1230.

Dahl, G. B., Løken, K. V., & Mogstad, M. (2014). Peer effects in program participation. *American Economic Review*, *104*(7), 2049–74.

De Giorgi, G., Pellizzari, M., & Redaelli, S. (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics*, *2*(2), 241–75.

De Paula, A. (2017). Econometrics of network models. In *Advances in economics and econometrics: Theory and applications: Eleventh world congress* (Vol. 1, pp. 268–323).

Dhaliwal, I., Duflo, E., Glennerster, R., & Tulloch, C. (2013). Comparative cost-effectiveness analysis to inform policy in developing countries: a general framework with applications for education. *Education policy in developing countries*, *17*, 285–338.

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

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, *101*(5), 1739–74.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from kenyan primary schools. *Journal of Public Economics*, *123*, 92–110.

Duflo, E., & Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly journal of economics*, *118*(3), 815–842.

Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of research in reading*.

Fruehwirth, J. C. (2013). Identifying peer achievement spillovers: Implications for desegregation and the achievement gap. *Quantitative Economics*, *4*(1), 85–124.

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2*(3), 205–27.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of applied econometrics*, *18*(5), 527–544.

Hanushek, E. A., & Woessmann, L. (2015). *Universal basic skills what countries stand to gain: What countries stand to gain*. OECD publishing.

Hesselius, P., Nilsson, J. P., & Johansson, P. (2009). Sick of your colleagues'absence? *Journal of the European Economic Association*, *7*(2-3), 583–594.

Hoxby. (2000). Peer effects in the classroom: Learning from gender and race variation. *National Bureau of Economic Research*.

Hoxby, C. M., & Weingarth, G. (2005). Taking race out of the equation: School reassignment and the structure of peer effects. *mimeo*.

Imberman, S. A., Kugler, A. D., & Sacerdote, B. I. (2012). Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *American Economic Review*, *102*(5), 2048–82.

Johnson, H., McNally, S., Rolfe, H., Ruiz-Valenzuela, J., Savage, R., Vousden, J., & Wood, C. (2019). Teaching assistants, computers and classroom management. *Labour Economics*, *58*, 21–36.

Lalive, R., & Cattaneo, M. A. (2009). Social interactions and schooling decisions. *The Review of Economics and Statistics*, *91*(3), 457–477.

Lavy, V., Paserman, M. D., & Schlosser, A. (2011). Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal*, *122*(559), 208–237.

Lavy, V., & Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, *3*(2), 1–33.

Lefgren, L. (2004). Educational peer effects and the chicago public schools. *Journal of urban Economics*, *56*(2), 169–191.

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

Lyle, D. S. (2007). Estimating and interpreting peer and role model effects from randomly assigned social groups at west point. *The Review of Economics and Statistics*, *89*(2), 289–299.

Machin, S., & McNally, S. (2008). The literacy hour. *Journal of Public Economics*, *92*(5-6), 1441–1462.

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

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, *60*(3), 531–542.

Marinelli, H. A., Berlinski, S., & Busso, M. (2021). Remedial education: Evidence from a sequence of experiments in colombia. *Journal of Human Resources*, 0320–10801R2.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), 415–444.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., ... others (2000). Manual for the patterns of adaptive learning scales. *Ann Arbor: University of Michigan*.

Ministerio de Educacion Nacional. (2016). *Derechos basicos de aprendizaje: Lenguaje*. Ministerio de Educacion Nacional.

Moffitt, R. A. (2001). Policy interventions, low-level equilibria, and social interactions. *Social dynamics*, *4*(45-82), 6–17.

Popova, A., Evans, D. K., Breeding, M. E., & Arancibia, V. (2021, 06). Teacher Professional Development around the World: The Gap between Evidence and Practice. *The World Bank Research Observer*.

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

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics*, *116*(2), 681–704.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the economics of education* (Vol. 3, pp. 249–277). Elsevier.

Stinebrickner, R., & Stinebrickner, T. R. (2006). What can be learned about peer effects using college roommates? evidence from new survey data and students from disadvantaged backgrounds. *Journal of public Economics*, *90*(8-9), 1435–1454.

Tincani, M. (2017). Heterogeneous peer effects and rank concerns: Theory and evidence, CESifo Working Paper Series. *CESifo Working Paper Series*.

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

Whitmore, D. (2005). Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. *American Economic Review*, *95*(2), 199–203.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and statistics*, *85*(1), 9–23.

Tables

Table 1: Baseline School and Individual Characteristics by Treatment Group

	Treatment schools		Control Schools		p-value Treatment = Control
	Means	S.D.	Mean	S.D.	
Panel A	*School and class characterisics*				
Class size	30.783	6.891	30.780	6.055	0.969
Eligible share	0.254	0.145	0.280	0.154	0.428
Panel B	*Individual characteristics - Low achieving students*				
Age	8.563	0.943	8.569	0.987	0.922
Gender (girl)	0.490	0.500	0.515	0.500	0.437
SES	0.298	0.458	0.319	0.466	0.488
Literacy score	93.068	23.625	92.517	26.328	0.773
Math score	20.981	7.694	20.306	8.048	0.195
Total score	114.049	27.119	112.823	29.724	0.587
Panel C	*Individual characteristics - Higher achieving students*				
Age	8.435	0.874	8.381	0.820	0.155
Gender (girl)	0.515	0.500	0.505	0.500	0.742
SES	0.233	0.423	0.219	0.413	0.497
Literacy score	139.376	28.057	140.447	28.371	0.616
Math score	25.757	8.072	25.487	7.782	0.508
Total score	165.133	30.751	165.935	30.929	0.713

Notes: Panel A: school characteristics. Panels B and C: individual characteristics. p-values are for tests of equality of the means across treatment and control groups. SES is an indicator for whether the child belongs to the bottom strata of the wealth distribution. This is based on the System of Identification of Social Program Beneficiaries (SISBEN) scores, the social stratification classification scale used in Colombia to target social programs.

Back

Table 2: Treatment Effects - Low Achieving Students

	(1)	(2)	(3)	(4)
Panel A: Literacy	0.362	0.361	0.358	0.342
	(0.091)	(0.091)	(0.091)	(0.078)
Panel B: Math	0.092	0.081	0.079	0.131
	(0.060)	(0.059)	(0.061)	(0.057)
Panel C: Total score	0.317	0.314	0.312	0.315
	(0.083)	(0.083)	(0.083)	(0.072)
Observations	1889	1889	1889	2413
Individual controls		✓	✓	✓
Class controls			✓	

Notes: The outcome variables are standardized test scores. Individual controls include a second-order polynomial in age and gender. Class control include class size and number of classrooms in the schools. All regressions control for school fixed effects. Robust standard errors are clustered at the classroom level, and presented in parentheses. Column 4 reports the results in the evaluation sample in Marinelli, Berlinski, & Busso (2021).

Back

Table 3: Reduced-form Estimates: Higher-achieving Students

	(1)	(2)	(3)
Panel A: Literacy	0.108	0.112	0.118
	(0.058)	(0.058)	(0.057)
Panel B: Math	0.034	0.035	0.038
	(0.049)	(0.049)	(0.049)
Panel C: Total score	0.112	0.116	0.120
	(0.056)	(0.056)	(0.055)
Observations	5181	5181	5181
Individual controls		✓	✓
Class controls			✓

Notes: The outcome variables are standardized test scores. Individual controls include a second-order polynomial in age and gender. Class controls include class size and number of classrooms in the schools. All regressions control for school fixed effects. Robust standard errors are clustered at the classroom level, and presented in parentheses.

Back

Table 4: Linear-in-means Model of Peer Effects in Academic Achievement

	Literacy			Total		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A - OLS						
Peers' average end-line score	0.535	0.533	0.527	0.560	0.559	0.555
	(0.048)	(0.048)	(0.049)	(0.046)	(0.046)	(0.046)
Panel B, 2SLS: first stage						
Treated class	0.159	0.159	0.168	0.159	0.159	0.166
	(0.059)	(0.059)	(0.058)	(0.056)	(0.056)	(0.055)
Panel C, 2SLS: second stage						
Peers' average end-line score	0.679	0.705	0.704	0.704	0.725	0.725
	(0.185)	(0.179)	(0.171)	(0.169)	(0.165)	(0.158)
Observations	5181	5181	5181	5181	5181	5181
Individual controls		✓	✓		✓	✓
Class controls			✓			✓

Notes: Estimates from models of dependent variable in column heading as a function of peers' average test scores. The outcome variables are standardized test scores. Individual controls include a second-order polynomial in age and gender. Class controls include class size and number of classrooms in the schools. All regressions control for school fixed effects. Robust standard errors are clustered at the classroom level, and presented in parentheses. Panel A reports the OLS results. Panel B and C report, respectively, the first-stage and second-stage results of a 2SLS model using random assignment to treatment as an instrument for peers' average end-line score.

Back

Table 5: Heterogeneity by Baseline Achievement

	Literacy			Total		
Panel A: Linear-in-means model						
Peers' average end-line score		0.704			0.725	
		(0.171)			(0.158)	
Quantile of baseline achievement	Q1	Q2	Q3	Q1	Q2	Q3
Panel B, 2SLS: first stage						
Treated class	0.217	0.125	0.183	0.199	0.141	0.169
	(0.062)	(0.056)	(0.067)	(0.061)	(0.058)	(0.062)
Panel C, 2SLS: second stage						
Peers' average end-line score	0.277	0.539	0.777	0.339	0.546	0.832
	(0.265)	(0.458)	(0.374)	(0.283)	(0.361)	(0.376)
Observations	1726	1726	1724	1748	1715	1712

Notes: Estimates from models of dependent variable in column heading as a function of peers' average test scores. The outcome variables are standardized test scores. Controls include a second-order polynomial in age and gender, class size and number of classrooms in the schools. All regressions control for school fixed effects. Robust standard errors are clustered at the classroom level, and presented in parentheses. Panel A reports the second stage of the 2SLS model in columns 3 and 6 of Table 4. Panels B and C report the first and second stages of a 2SLS model using random assignment to treatment as an instrument for peers' average end-line score separately for children in the first, second, and third terciles of the baseline achievement distribution of the outcome variable.

Back

Table 6: Linear-in-means Model of Peer Effects in Academic Achievement - Treatment Schools Only

	Literacy	Total
	(1)	(2)
Panel A: OLS		
Peers' average end-line score	0.44	0.501
	(0.1)	(0.086)
Panel B, 2SLS: reduced form		
Share of eligible students	0.789	0.609
	(0.328)	(0.357)
Panel C, 2SLS: first stage		
Share of eligible students	0.77	0.485
	(0.251)	(0.262)
Panel D, 2SLS: second stage		
Peers' average end-line score	1.025	1.255
	(0.328)	(0.553)
Observations	2602	2602

Notes: The outcome variables are standardized test scores. Controls include a second-order polynomial in age, gender, peers' baseline average achievement and school fixed effects. Robust standard errors are clustered at the classroom level, and presented in parentheses. Panel A reports the OLS results. Panel B reports the reduced form. Panel C reports the first stage, and Panel D reports the second stage of a 2SLS model using the share of treated students as an instrument for average end-line test scores.

Back

Table 7: Teachers Reports

	PALS			
	Teaching Efficacy	Performance Approaches	Student bad-behaviour	Phonics approach
	(1)	(2)	(3)	(4)
Treated class	0.081	0.039	-0.367	0.029
	(0.273)	(0.215)	(0.293)	(0.348)
Control mean†	0	0	0	0
Observations	70	70	70	45

Notes: The outcome variables in columns 1 to 3 are factor scores constructed from the Patterns of Adaptive Learning Scales (PALS). The outcome variables in column 4 is a factor score constructed from questions inquiring directly about the "Phonics approach" used in treatment schools (e.g. "Phonological awareness is indispensable for reading"). Controls include teacher age, gender and experience. Standard errors are presented in parentheses. † The scales have been standardized to have a mean of zero and a standard deviation of one in the control group.

Back

Table 8: Peer-to-peer Interactions - Treatment Schools Only

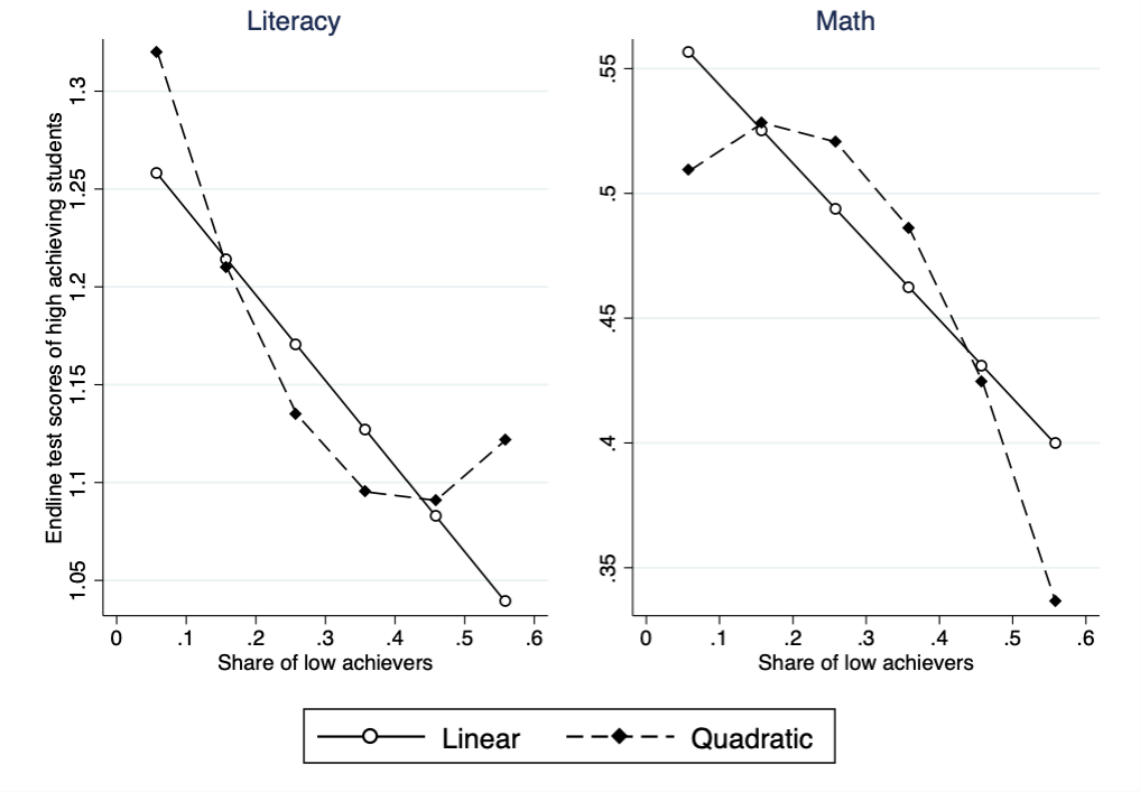
	Literacy (1)	Total (2)
Panel A: OLS		
Δ_c^E	0.004	0.003
	(0.002)	(0.002)
Panel B, 2SLS: reduced form		
Average attendance	0.538	0.488
	(0.235)	(0.226)
Panel C, 2SLS: first stage		
Average attendance	13.97	10.794
	(2.354)	(2.423)
Panel D, 2SLS: second stage		
Δ_c^E	0.039	0.045
	(0.018)	(0.023)
Observations	2602	2602

Notes: The outcome variables are standardized test scores. Δ_c^E is defined as the average change in raw test scores between baseline and endline for low-achieving children in the class. Controls include a second-order polynomial in age, gender and school fixed effects. Robust standard errors are clustered at the classroom level, and presented in parentheses. Panel A reports the OLS results. Panel B reports the reduced form. Panel C reports the first stage, and Panel D reports the second stage of a 2SLS model using average tutorial attendance as an instrument.

Back

Figures

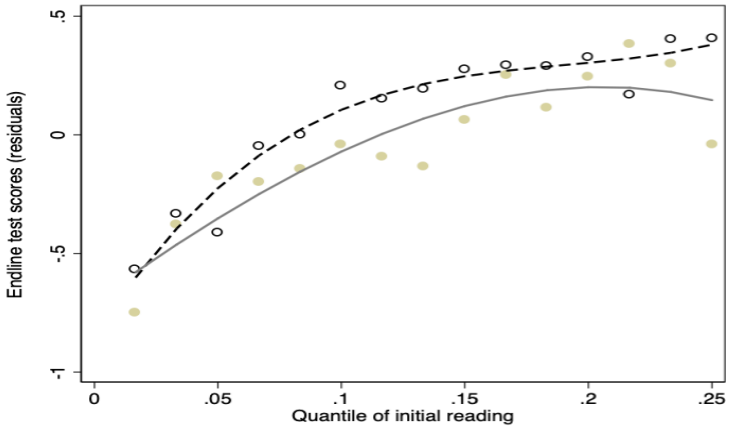
Figure 1: Linear and Quadratic Fits of End-line Scores of Higher-achieving Students by Classroom Share of Low Achievers (Control Schools Only)



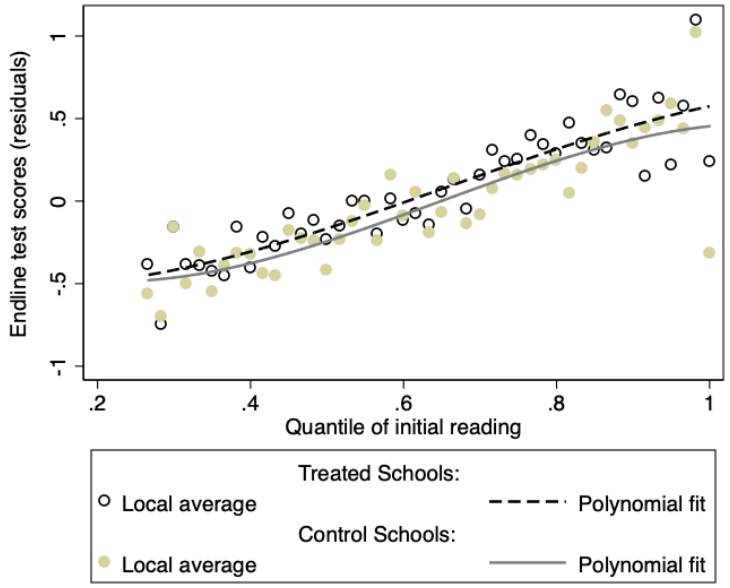
Lines represent linear and quadratic fits of standardized end-line test scores of high-achieving students as a function of baseline share of low achievers in non-treatment schools. Controls include a second-order polynomial in age, gender and school fixed effects. The figure is trimmed at the 5th and 95th percentiles of the distribution of classroom share of low achievers.

Back

Figure 2: Local Averages and Polynomial Fits of End-line Scores by Quantile of Baseline Reading



(a) Low-achieving Students

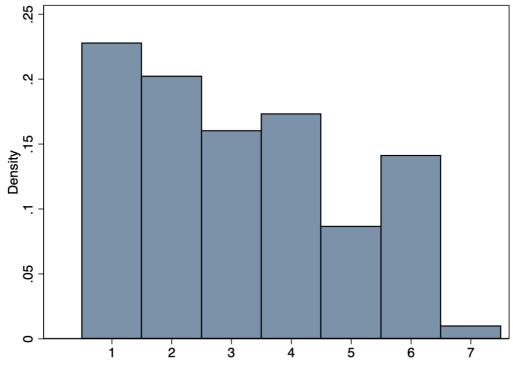


(b) Higher-achieving Students

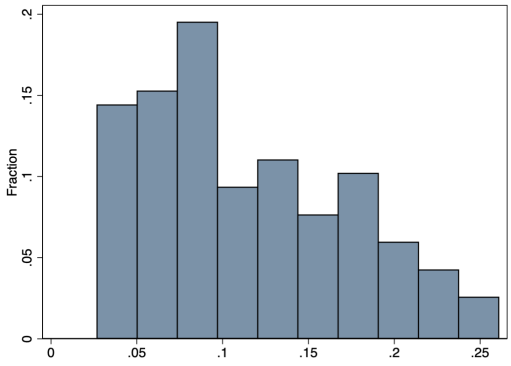
Dots represent local averages. Lines represent polynomial fits of end-line test scores as a function of baseline scores estimated using a second-order polynomial. The variable used to construct these figure is the residuals of standardized test scores obtained from a regression of end-line test scores on a second-order polynomial in age, gender and school fixed effects estimated separately for low-achieving students (Panel A) and high-achieving students (Panel B). By construction, the residuals are centered around zero.

Back

Figure 3: Peers in the Same Tutorial Group and Class Size Reduction (Treatment Schools Only)



(a) Number of Same Class Students per Group



(b) Class Size Reduction (%)

Panel A shows the number of low-achieving students that attend the same class and were assigned to the same tutorial group. Panel B shows the class size reduction in our sample expressed in terms of total number of students in the classroom. The average class size reduction is three. In the sample 75 percent of classes experienced a reduction of 4.5 or fewer students.

Back

Appendix

Appendix Tables and Figures

Table A1: Heterogeneity by Baseline Achievement - Interacted Model

	Literacy			Total		
	Q1	Q2	Q3	Q1	Q2	Q3
Panel A: First stage						
Treated class	0.655	0.68	0.754	0.527	0.546	0.621
	(0.044)	(0.049)	(0.062)	(0.042)	(0.035)	(0.058)
Panel B: Second stage						
Peers' average endline score	0.069	0.614	1.148	-0.301	0.485	1.119
	(0.253)	(0.252)	(0.245)	(0.296)	(0.293)	(0.276)
Panel C: F-test of equality (p-value)						
H0: $\rho_1 = \rho_2$		0.000			0.000	
H0: $\rho_1 = \rho_3$		0.000			0.000	
H0: $\rho_2 = \rho_3$		0.000			0.000	
Observations		5181			5181	

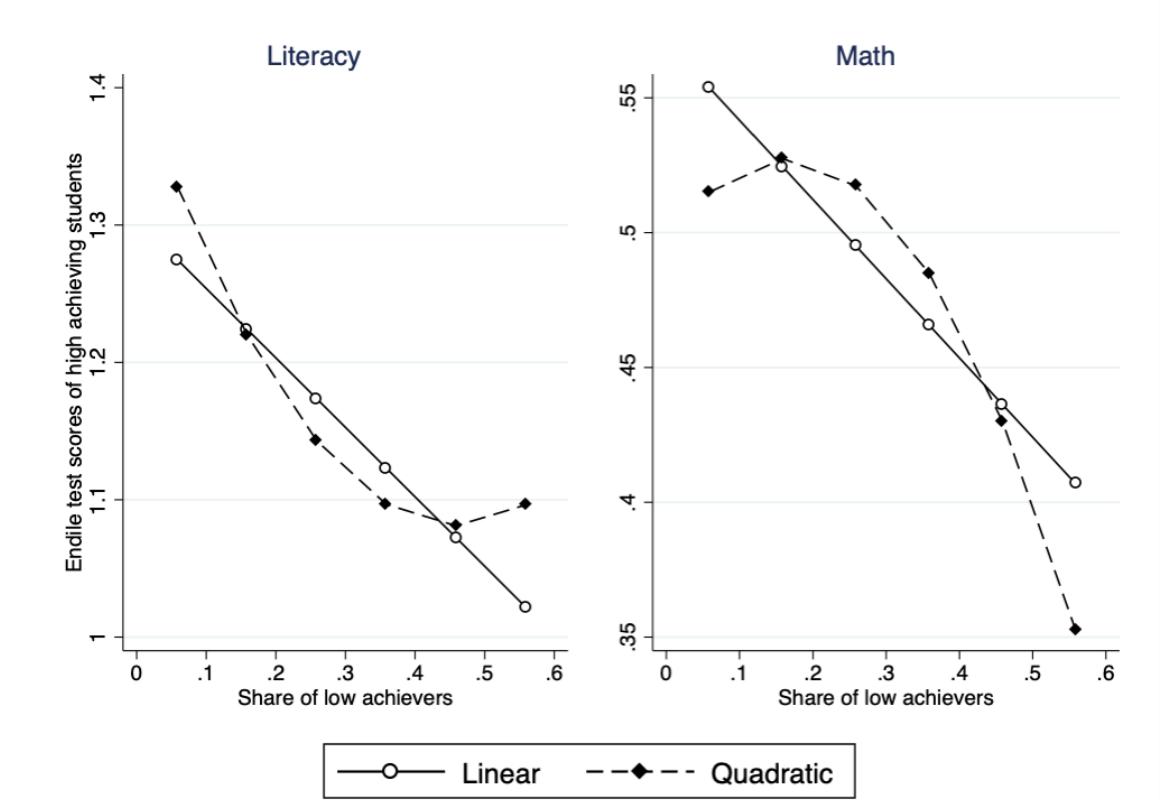
Notes: Estimates from models of dependent variable in column heading as a function of peers' average test scores. The outcome variables are standardized test scores. Controls include a second order polynomial in age and gender, class size and number of classrooms in the schools. All regressions control for school fixed effects. Robust standard errors are clustered at the classroom level and presented in parenthesis. Panels A and B report the first and second stages of a 2SLS model using classroom random assignment to treatment interacted with student quantiles as an instrument for peers' average endline score.

Table A2: Reduced Form Impacts on Higher Achievers by Class Size Reduction

	(1)	(2)	(3)	(4)	(5)	(6)
	Literacy		Math		Total score	
	Low†	High*	Low	High	Low	High
Treated class	0.131	0.084	-0.021	0.043	0.109	0.094
	(0.085)	(0.072)	(0.064)	(0.061)	(0.082)	(0.070)
Observations	2597	2583	2597	2583	2597	2583

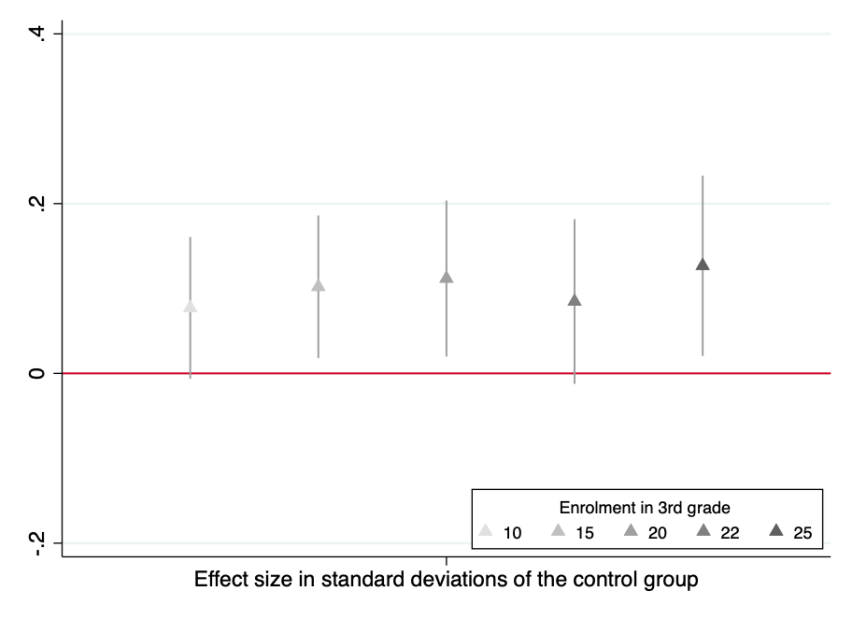
Notes: The outcome variables are standardized test scores. Controls include a second order polynomial in age and gender, class size and number of classrooms in the schools. All regressions control for school fixed effects. Robust standard errors are clustered at the classroom level and presented in parenthesis. † indicates that the class experienced a small reductions in size (below the median of 3.5 students). * indicates that the class experienced a large reductions in size (above the median of 3.5 students).

Figure A1: Linear and Quadratic Fits of End-line Scores of Higher-achieving Students by Classroom Share of Low Achievers (Control Schools Only)



Lines represent linear and quadratic fits of standardized end-line test scores of high-achieving students as a function of baseline share of low achievers in non-treatment schools. Controls include a second-order polynomial in age, gender, school fixed effects and the peers' SES. The figure is trimmed at the 5th and 95th percentiles of the distribution of classroom share of low achievers.

Figure A2: Reduced-form Estimates on Higher-achieving Student - Robustness



The figure shots the point estimates and 95% confidence intervals of the reduced-form effects on higher-achieving students for different samples defined on the basis of enrolment in third grade.