Behavioral inattention and human capital accumulation*

Manuel Barron

September 2019

Abstract

I use data from a standardized test applied to second and eighth graders in rural Peru to show that inability to correctly interpret test scores can affect schooling outcomes persistently. Marginally classifying as "remedial" in second grade math reduces rural males' eighth grade scores by 0.18 standard deviations, compared to students that obtained marginally higher scores and were classified as "in transition" in second grade. This paper provides novel evidence on the mechanisms at play, showing that results owe to classroom environment and household resource reallocation.

JEL Codes: D12, D83, D91, J24

Keywords: Behavioral Inattention, Human Capital, Educational Investment

*Department of Economics, Universidad del Pacifico MF.BarronA@up.edu.pe. I am indebted to Pia Basurto, Juan Castro, and participants at the 6th Annual Meeting of the Peruvian Economic Association for insightful comments and Lucina Rodriguez for superb research assistance. Special thanks to the Ministry of Education's Office of Monitoring and Strategic Evaluation and UMC for sharing the anonymized ECE data and their invaluable feedback. All errors are mine.

1 Introduction

Behavioral inattention, also known as limited attention or incomplete attention, is a cognitive bias that consists of fixating on certain attributes of a good or quantity, neglecting the rest at least to a degree. This phenomenon has been documented in a wide array of settings, from purchases of beer, to cars, health plans and mortgages, but its consequences have not been fully explored yet. In this paper, I study the consequences of behavioral attention to test scores on children's human capital accumulation. This is a particularly interesting setting to study behavioral inattention, because children, as well as their teachers and parents have a wealth of information about the students abilities and skills besides test scores.

The study setting is Peru, a developing country with an educational system that faces challenges common to the developing world. Despite access to education is broad, learning outcomes are precarious.[1] All second and eighth graders in the country take the Census Student Evaluation (*Evaluación Censal de Estudiantes*, in Spanish, or *ECE*), a standardized test that evaluates math and language skills. A few months after the test, students and their parents receive their test score and a label based on their performance, which in second grade can be "remedial", "in transition", or "proficient". Two students with essentially identical skills can attain different labels if their scores fall at different sides of the threshold. Under full attention, observing their scores would lead to conclude that these students had indeed similar performance. However, behavioral inattention could lead the students, their parents, teachers or peers, to focus on the label instead of the score; so these students would be perceived differently just because they received different labels.

Labels have long been studied in sociology and education (Becker, 1963; Rist, 1977). A small, recent literature estimates empirically their causal effects on learning outcomes. For instance, Avery et al. (2018) shows that marginally earning higher integer

[1] In the 2015 Program for International Student Assessment (PISA) evaluation, which evaluated skills in science, reading, and mathematics, 46% of Peruvian students were ranked as low achievers in its three subjects, and only 0.6% was ranked as top performers in at least one subject. The OECD averages were 15% (top performers) and 13% (low achievers). (OECD, 2018).

score in advanced placement exams in a given subject increases the probability that students choose said subject as a major, even when the test does not count towards college credit. Smith, Hurwitz, and Avery (2017) shows that this also increases the likelihood college completion. Thus, labels can lead to resource misallocation. An important feature of these studies is that the exact scores were not revealed to the students. Therefore, in principle, revealing a student's exact score could make the effects disappear. However, Papay, Murnane, and Willett (2016) shows that labels may have an effect even when the students and their parents receive their exact scores in addition to the label, and this is likely due to behavioral inattention. Using a similar identification strategy as this study, the authors show that low-income students in Massachussets who marginally earn a positive label in a low-stakes math test in high school are more likely to attend college. To my knowledge, this is the only previous evidence on the effects of inattention to test scores. My study builds on this evidence to provide two contributions to the literature of empirical behavioral economics: (i) it provides evidence of inattention to test scores in a developing country context, and (ii) it offers novel evidence on the mechanisms behind these effects.

First, I use a regression discontinuity design to show that the labels obtained in second grade math causally affect performance six years later among rural males. Math performance is of special interest because it has been shown to be a good predictor of readiness for post-secondary education and future labor market outcomes (Altonji and Blank, 1999; Card and Payne, 2017). Marginally classifying as "remedial" in second grade math reduces the overall score by 0.18 standard deviations in eighth grade, with reductions in both math and language (of 0.11 and 0.21 standard deviations, respectively). These magnitudes are comparable to the average effects of interventions of structured pedagogy, which are the most effective type of educational intervention found by Snilstveit et al. (2015), showing how important can be behavioral inattention in the process of human capital accumulation. Furthermore, barely classifying as "remedial" in second grade math increased the probability of being classified as "warning" in eighth grade math and language. In line with Papay, Murnane, and Willett (2016),

these results suggest that labels, even from low-stakes tests, can affect educational outcomes permanently, especially among vulnerable groups.

Second, I investigate the mechanisms behind these effects, which is the main contribution of this study over the previous empirical evidence. To do so I exploit information on socioeconomic background, household investment in educational inputs, child labor, perception of own academic skills, interaction with parents at home, and the student's perception of the classroom environment. The analysis reveals that scoring marginally below the remedial cutoff in second grade reduces household investment in educational inputs, increases child labor, and makes students more likely to perceive a negative classroom environment. The effects on perceptions of own skills and on an index of family support are negative, but imprecisely estimated, suggesting the first three channels are more important in this setting.

The dataset has three main caveats worth considering. First, second-grade performance data was matched for roughly 55% of all eighth graders that took the test in 2015 (and 35% of rural males). Overall, unmatched students belong to a lower socioeconomic status than matched students. However, section 5.4 shows that there are no signs that this threatens the internal validity of the findings reported here. Second, by its nature, test score data is available only for students who had not dropped out of school by eighth grade. Moreover, since this is the first cohort who took the test in both grades, these are students who had not been held back between second and eighth grade. If behavioral inattention caused students to fail a grade or drop out of school, the estimates reported here would be lower bounds on the total effect of inattention on learning. A third, and perhaps the most important, caveat relates to the timing of the mechanisms. This study uncovers changes that have persisted through time, which makes them interesting and important, but the data does not allow to disentangle the timing of these effects. It does not allow show whether the changes in child labor, book ownership and perceived school environment were immediate consequences of the test outcomes or were caused by another variable.

The paper also speaks to other strands of literature. First, to the literature on

human capital investment, showing that human capital investment can be affected by cognitive biases. Second, to the labeling literature, providing evidence on the unintended effects of labeling at early ages. Third, it contributes to the debate on the effects of child labor. While some studies find no or weak substitutability between child labor and schooling outcomes (Ravallion and Wodon, 2000; Patrinos and Psacharopoulos, 1997), the evidence presented here suggests that child labor is associated with lower learning even among children that are making adequate progress in school. Baland and Robinson (2000) argue that child labor may materialize if parents fail to fully internalize its negative effects. This study shows that it may also happen if the students academic skills are underestimated.

The evidence provided by this study has two clear policy implications. First, the school should foster improvements in classroom environment. Second, communication of test scores must be improved. The school should also provide guidance and tools to parents and teachers to help children who are labeled as remedial. This is especially important given the higher likelihood of receiving a "warning" label in eighth grade which could lead parents, teachers and peers to reinforce the behaviors and attitudes triggered by the second grade test results leading to a self-fulfilled prophecy.

The paper is organized as follows. The next section describes the theoretical background, linking behavioral inattention with learning outcomes and fleshing out the mechanisms that relate these variables. Section 3 describes the data and the study setting. Section 4 details the empirical approach. Section 5 presents the results and discusses its robustness, and section 6 concludes.

2 Theoretical background

When making choices, our attention is limited, because of the large number of dimensions to be considered in each choice and because of the high frequency with which we need to make choices. In practice, we forcefully take into account only a few considerations, usually those that are most salient to us. Neglecting the rest may carry serious

consequences. For instance, a branch of literature has shown that consumers fail to assess the full costs of purchasing a good.[2] Most notably, Chetty, Looney, and Kroft (2009) and Taubinsky and Rees-Jones (2017) show inattention to taxes when they are not included in the tag price. In turn, Ellison (2005); Gabaix and Laibson (2006); Ellison and Ellison (2009), and Brown, Hossain, and Morgan (2010) show inattention to other *shrouded attributes*, like small-font bank fees, or "shipping and handling" costs. Inattention has also been documented in automobile purchases (Lacetera, Pope, and Sydnor, 2012), health plan choices (Abaluck and Adams, 2017), and failing to refinance mortgages (Andersen et al., 2015). Some studies, on the other hand, find results consistent with full attention to key attributes of large purchases. For instance it has been documented that consumers in the US pay full attention to fuel prices when considering the purchase of a home (Myers, 2018) or cars (Busse, Knittel, and Zettelmeyer, 2013).

Incomplete attention could play a role in education, especially in how students, parents and teachers use test outcomes to estimate a student's academic skills. Test scores usually have labels attached to them, like "remedial" or "proficient", that help interpret the score obtained by the student. Scores contain finer information about performance than labels, but are more difficult to interpret. Behavioral inattention to scores would lead to systematic biases in the estimation of student skills, and it could appear if individuals focus on the labels and disregard the scores.

To my knowledge, the only study that has shown evidence of behavioral inattention to scores is Papay, Murnane, and Willett (2016). The authors show that marginally earning positive labels in low-stakes standardized tests during high school increases college enrollment compared to students that scored barely below those thresholds, despite both the labels and the exact scores being known by the students and their parents.[3] Papay, Murnane, and Willett (2016) argue that performance in tests can

[2]Gabaix (2019); DellaVigna (2009), and Caplin (2016) provide more comprehensive literature surveys on behavioral inattention and theoretical tools to model this cognitive bias.

[3]Two recent studies shed light on the causal effect of labels by showing that marginally receiving a higher integer score in high-stakes advanced placement exams affects college outcomes and choice of majors in the US (Smith, Hurwitz, and Avery, 2017; Avery et al., 2018). However, these are not inattention studies because the continuous scores are known only to the econometrician, not to the students, their families or teachers. A relatively larger, but less related, literature studies the consequences of barely failing examinations (Machin,

affect students by making them eligible for educational interventions, by affecting the student's self-judgments, or by influencing the perceptions of parents, teachers, and peers. The first mechanism is not relevant in this setting because no interventions or policies are determined by the student's individual test results. Thus, I investigate whether test scores affect perceptions about own skills, as well as the perceptions of parents, teachers and peers. If a student is inattentive to scores, her perceptions about own skills will be directly affected by the label attached to her. Two comparable students with similar scores that happen to fall at different sides of a given threshold will obtain different labels, which will lead them to perceive themselves differently. Perceiving oneself as less skilled can affect motivation, time studying, and overall investment in education, hindering future performance and thus leading to self-fulfilled prophecies. The same reasoning applies to peers, teachers and parents, who may treat a student differently based on her perceived skills.

Standard economic models of educational investment suggest that households should reallocate resources in response to a change in expected returns to schooling arising from low test scores. Dizon-Ross (2019) provides evidence of this behavior in Malawi. Behavioral inattention implies that households will overreact to more salient information, namely the label attained by the student, and neglect information that is more complicated to process or remember, like the test score itself. Under both the standard and the behavioral framework, facing a lower return to formal education, parents may purchase fewer inputs of academic skill formation, like books, and focus on fostering non-academic skills that can counteract the (perceived) lack of academic skills once their child enters the labor market.[4] The difference is that, if parents are fully attentive, the reallocation would be smooth across the threshold, whereas if parents exhibit behavioral inattention the reallocation would exhibit a discontinuity at the threshold.

Similarly, if teachers or peers are inattentive to scores, they may change their atti-

McNally, and Ruiz-Valenzuela, 2018; Papay, Murnane, and Willett, 2010, 2014; Reardon et al., 2010).

[4]In rural settings, farm labor is the main substitute of formal schooling in income generation (Jacoby, 1994; Glewwe and Jacoby, 1994). In fact, Meza and Pérez (2018) find that child labor in rural Peru is negatively associated with schooling attainment but not with income during adulthood, suggesting that additional experience may indeed compensate for reduced formal schooling in this setting.

tudes towards students that marginally classify as remedial. Teachers may lower their expectations of student performance, which has been shown to affect student outcomes (Jussim and Harber, 2005; Rosenthal and Jacobson, 1968; Cooper and Good, 1983; Spencer, Steele, and Quinn, 1999; Papageorge, Gershenson, and Kang, 2016). More recently, Carlana (2019) and Alesina et al. (2018) show that teacher bias against girls and immigrants, respectively, reduce learning outcomes against the stigmatized groups. The effects of negative teacher attitudes could be worsened by the interaction with peers. For instance, Kristoffersen et al. (2015) studies the detrimental effects of disruptive students on their peers academic achievement in Denmark. A larger literature studies the detrimental effects of bullying on learning (Brown and Taylor, 2008; Eriksen, Nielsen, and Simonsen, 2014; Gutierrez, Molina, and Ñopo, 2018). Interaction with peers is especially relevant in Peru, since 71% of children aged 9-11 had suffered psychological violence from their peers at some point in their lifetime, while 40% had suffered physical violence (75% had suffered either) (INEI, 2016a). Conversely, an improved classroom environment can lead to academic gains, which are mediated through lower levels of classroom disruption and violence, improved inter-student and student-teacher relationships, and lessened teachers' fatigue (Lavy and Schlosser, 2011).

3 Data and study setting

The data source used in this study is the Student Census Evaluation, *ECE* for its name in Spanish. *ECE* is a standardized test that the Ministry of Education applies annually to measure learning outcomes among second graders since 2006. The test constitutes a milestone in school dynamics, especially in primary-level public schools, and its results are awaited by the school community and are an important determinant of institutional climate (Sempé et al., 2017). Teachers hold meetiings with parents, where they explain the results and the labels. Since 2015 the Ministry applies it to eighth graders as well, which is the second year of high school in the Peruvian educational system. The evaluation is applied to all students in the respective grade

in all private and public schools across the country. Since 2015, the school-level results partly determine the allocation of *Bono Escuela*, a monetary bonus allocated to the school teaching staff, but from the student's perspective *ECE* is a low-stakes test. The second grade evaluation tests the students in two subjects: math and language. The eighth grade evaluation tests the students in the same subjects.[5] Based on their score in each component, students can be classified as remedial ("en inicio"), in transition ("en proceso"), or proficient ("satisfactorio"). The eighth grade test includes an additional category, lower than remedial: "pre-inicio", which I translate as "warning" to match labels used in other studies. Students and their parents receive both the child's score and the label attached to it. However, despite *ECE* results are delivered to parents, the ministry's recommendations are not translated to teaching practices (Sempé et al., 2017).

The analysis in this paper is performed with data from students who took the test in 2009 as second graders and in 2015 as eighth graders, the first cohort to take the test in both grades. Importantly, the eighth grade evaluation includes a post-test survey that allows to investigate the mediating channels discussed in section 2. There are sets of questions on perceptions about family support (e.g. "I talk with my parents about school work"), own perception of academic skills (e.g. "I can understand difficult topics in math") and perception of classroom environment (e.g. "in my school, teachers are respectful of our answers even when we are wrong"). There are 20 questions on student perceived academic skills, 15 questions on interaction with their parents and family, and 21 questions on climate in the classroom. To avoid false positives that could arise from analyzing the questions individually, I used principal component analysis (PCA) to create indices with all the available questions under each category. Some statements denote desirable outcomes, like "my teachers give us additional help when we need it", while others refer to undesirable outcomes, like "my teachers make me feel bad when I make a mistake". To construct each index I created a set of indicators, one for

[5] In 2016 a third subject was incorporated "social sciences" which includes history, geography and economics.

each statement, that take the value of 1 if the student replied "frequently" or "always" to a desirable outcome, or "rarely" or "never" to an undesirable outcome. PCA was conducted on these indicators, and the index was constructed with the first principal component. The list of variables in each index, as well as the mapping to the indicators used for the PCA, are reported in Tables A.11 through A.13.

The main descriptive statistics for the sample of students with data for their second and eighth grade tests are presented in Table 1. The sample is split by student gender and school location (urban or rural). The definition of rurality used by the Ministry of Education is stricter than the more commonly used definition of having fewer than 2,000 people in a settlement, mainly because the Ministry's definition pools together the population of settlements that are geographically contiguous. Table 1 shows small and non-systematic differences between males and females in socioeconomic characteristics and learning outcomes, but rural students are notably worse off than their urban counterparts. Parental education is lower in rural areas, with only 18% of mothers and 30% of fathers having completed secondary schooling, compared to more than 60 and 70% in urban areas, respectively. Rural students are also more likely to have a parent who speaks an indigenous language than urban students, at 24 and 4%, respectively. The Ministry constructed a socioeconomic index based on asset ownership, which also indicates sizable differences across populations. On average, rural students have a socioeconomic index 1.4 standard deviations lower than that of urban students. These differences are reflected in school performance. In second grade *ECE*, 55% of rural students were classified as remedial in math and 33% in language; while the figures for urban students were 32 and 10%, respectively. The eighth grade *ECE* shows similar differences: 58% of rural students were classified as "warning" in math and 48% in language, compared to 27 and 14%, respectively, of urban students.[6]

The mechanisms suggested by the literature that are relevant in this setting, as

[6]Descriptive statistics for unmatched students, those for whom there is no data on their second grade test, are provided in appendix table A.1. Overall, matched students are better-off in socioeconomic status and test performance. Section 5.4 shows that this type of sample selection is unlikely to affect internal validity, as all available observable characteristics are balanced at the threshold.

discussed in section 2 are family support, classroom environment, perceptions of own skills, and household investment in education. Table 2 shows that there are considerable differences across genders in these variables. Males have weaker interaction with their families than females. Somewhat surprisingly, males report lower confidence on their own academic skills than females, with a difference of 0.10 standard deviations in rural areas and 0.15 in urban areas. On average, males have a more negative perception of their classroom environment, at 0.15 standard deviations in rural areas and 0.11 standard deviations in rural areas. There is a clear difference by area in this dimension, as the index for rural males is 0.16 standard deviations higher than that of their urban counterparts. Book availability is similar across genders, but as expected, books are more widely available in urban areas. Students are also asked if they would drop out immediately if they could. Students who say they would drop out are asked the reason, and 41% of rural males report "because I have to work", which is the proxy I use for child labor. The figure for rural females is similar, at 38%, and lower in urban areas, at 18 and 11% for males and females, respectively. Regrettably, this question is not asked to all students. However, this figures are well in line with country-wide child labor figures: in rural Peru 52% of children and youths aged 5-17 work, while the figure for urban areas is 16% (INEI, 2016b).

4 Empirical approach

I apply a regression discontinuity design around the threshold used by the Ministry of Education for classifying students as "remedial" versus "in progress".[7] The estimating regression is:

$$y_i = \alpha + f(x_i) + \delta I(x_i \leq c) + \varepsilon_i \tag{1}$$

y_i is student i's eighth grade score (overall, math or communications), x_i is the

[7]There is another category, "proficient", for scores above 640 but the RD design lacks statistical power to detect effects at this threshold. Appendix A.14 shows that, based on ex-post power calculations, the minimum detectable effects at these threshold are substantially larger than the effects observed in the data.

student's second grade math score, c is the threshold between "remedial" and "in transition" (512 points) and $f()$ is an unknown function, which is estimated with first-order local polynomials, using second-order polynomial to correct for bias, as suggested by Calonico, Cattaneo, and Titiunik (2014) and Calonico, Cattaneo, and Farrell (2018). In this equation, δ measures the effect of marginally scoring as remedial. A statistically significant value is evidence of behavioral inattention, while lack of significance is consistent with either complete attention or with that crossing the cutoff is not perceived as relevant. Standard errors are calculated using the heteroskedasticity-robust nearest neighbor variance estimator with at least three neighbors. Given that the treatment is the information received by the student, it is not necessary to cluster the standard errors (Abadie et al., 2017).[8]

Given the structural differences by area of residence and gender discussed in the previous section, I estimate the regressions separately for each subgroup. Since the main regression was estimated in four subsamples, I rely on the false discovery rate (FDR) q-values (Anderson, 2008) to avoid false positives. In addition, following the methodology outlined in Lee, Miguel, and Wolfram (2019), I first estimate a regression with an overall outcome as dependent variable, namely the total test score. If there is a significant change in the overall outcome, the regression is estimated also on its components, i.e. the math and language score (or on the reaching the different labels in each of these components), as well as on the mechanisms that should lead to these changes. Here again, I control for multiple hypotheses testing with FDR q-values. Regressions related to test components and mechanisms for which the overall score did not show significant changes are reported in the online appendix.

[8]Figure A.1 shows that the main results and their statistical significance are unchanged if the standard errors are calculated clustering at the province level or clustering at the province level and using three nearest neighbors.

5 Results

5.1 Effects on test performance

Figure 1 presents the main results in the paper. The horizontal axis measures second grade test scores, and the vertical axis measures eighth grade scores. Marginally scoring as "remedial" in second grade math significantly reduced eighth grade scores by 0.18 standard deviations among rural males (FDR q-value = 0.04). The optimal bandwidth is 38 points at each side of the threshold. Given the distribution of scores, this implies that the RD results are valid for 26% of rural males in the sample. Figure A.2 shows that the effects in the other three subsamples are small and not statistically significant. The effect being present only among rural males is consistent with males being expected to be better at math than females (Spencer, Steele, and Quinn, 1999) together with the results from Sempé et al. (2017) who find that ECE scores are much more important for students, parents and teachers in rural than urban schools, as discussed earlier.

Column 1 in Table 3 reports the effects of scoring just below the remedial cutoff in second grade on eighth grade scores, for the sample of rural males. Row 1 shows the effect on the overall score reported above, and rows 2 and 3 show the effects disaggregated by test component, math and language. By eighth grade, rural males lost 0.11 standard deviations in math and 0.21 standard deviations in language for having marginally classified as remedial in second grade math. These effects of similar magnitude as the average effect of structured pedagogy interventions, the most effective interventions in terms of learning found in the review by Snilstveit et al. (2015), and are larger than the average effects found in computer and community-based interventions (Snilstveit et al., 2015).

Columns 2-4 in Table 3 report the effects of second grade math score on the label attained in eighth grade, shedding further light on the nature of the effects found in Column 1. Scoring marginally below the "remedial" cutoff in second grade math increased the probability of classifying as "warning" in language six years later by 18 percentage points, respectively. More importantly, it increased the probability of

receiving a label of "warning" in both subjects by 14 percentage points. These are considerable effects, ranging from 36-37% of the overall rates. There are no statistically or economically significant effects on the probability of reaching other labels. Overall, this evidence suggests that marginally earning the "remedial" label in second grade math traps rural boys in the lowest end of the scores distribution.

5.2 Mechanisms

Table 4 investigates the mechanisms behind the effects reported in the previous section, grouped as interaction with family, perceptions of own skills, perception of classroom environment, and household investment in human capital. Marginally classifying as remedial in second grade math reduced the family support index by 0.12 standard deviations six years later, although the estimate is not statistically significant. Similarly, own perception of academic skills decreased by 0.15 standard deviations, but this coefficient is not statistically significant either. On the other hand, column 3 shows that the negative classroom environment index, which reflects the student's perception about teacher and peer behavior, increased by 0.29 standard deviations. Given the evidence on the effect of teacher and peer behavior on school performance, an effect of this magnitude could be instrumental in explaining to a considerable extent the findings reported in the previous section. Next, columns 4 and 5 explore mechanisms related to parental decisions. Marginally scoring as remedial in second grade math increases the probability of students reporting they have to work by 21 percentage points and reduces by 46 percentage points less likely to have five or more books available at home. Together, these columns suggest that parents invest significantly less in inputs like textbooks and increase child labor relative to parents of essentially identical children who barely escaped the "remedial" label in second grade math. This is in line with the main findings by Dizon-Ross (2019), the key difference being that parents in this study setting overreact to negligible differences in scores.

Thus, classroom environment and household resource reallocation are two likely mediating channels behind the effects caused by marginally attaining the remedial label

in second grade math. However, given that the data on outcomes and mechanisms was collected at the same time, it is not possible to assert whether these channels are a direct consequence of inattention to scores, if one of them triggered the other, or if they were both triggered by some other variable that is not observable in the dataset.

5.3 Robustness checks

This section discusses the results of two types of robustness checks: to changes in bandwidth and in the degree of the polynomial used to estimate the RD and to correct for the bias discussed in Calonico, Cattaneo, and Titiunik (2014). The results of the first robustness check are reported in Figure 2. The dependent variable in each panel are the standardized total, math, and language scores, the main outcome variables. The horizontal axis measures alternative bandwidths as a proportion of the optimal bandwidth selected by the methodology developed by Calonico, Cattaneo, and Titiunik (2014). The vertical axis shows the RD estimate obtained using the bandwidth indicated in the horizontal axis. The dotted lines indicate the RD estimate with the optimal bandwidth, as well as its 95% bias-corrected confidence interval. Panel (a) shows that in all cases the resulting coefficient on total score always falls within the 95% confidence interval of the optimal-bandwidth coefficient. Panel (b) shows the same for math score. Panel (c) shows that this is the case also for language score for all but the 50% bandwidth. Next, Figure 3 explores the robustness of the mechanisms that were found relevant in section 5.2. Panel (a) shows that the effects on the index of perceived classroom environment fall within the optimal-bandwidth confidence interval. Panels (b) and (c) show that the effects on family investment are less robust. The effects on book ownership are robust for bandwidths from 70 to 140% of the optimal bandwidth, while the effects on work are robust for bandwidths from 50 to 140% of the optimal bandwidth.

Figure 4 examines the sensitivity of the results to the functional form used to estimate equation (1). The figure compares the main results, obtained with local linear regressions, to the results obtained using local quadratic and local kernel regressions,

as suggested by Gelman and Imbens (2019). The point estimates are essentially unchanged, and fall well within the 95% confidence intervals of the benchmark model.

Therefore altogether, the effects on the main outcomes are robust to choice of bandwidth. The most robust mechanism is classroom environment, although book availability and child labor are robust over a considerable range of bandwidths. The effects on outcomes and mechanisms are robust to the functional form used to estimate equation (1).

5.4 Threats to validity

Regression discontinuity design relies on two main identification assumptions: that no other variable must change discontinuously at the threshold and that there is no manipulation of the running variable at the threshold. The dataset contains a number of variables that can assist in assessing the first assumption, including socioeconomic status, parental education, parental ethnicity, and probability of having attended preschool. Using the same specification as equation (1), Table A.2 shows that none of these characteristics varies discontinuously at the "remedial" threshold.

The second assumption requires no manipulation of the scores at the threshold, which could indicate differences in unobservable variables at the threshold. Manipulation in this context is highly unlikely given that each student takes the test only once, they are not aware of the correct answers and they do not know the score they will obtain, so there is no realistic way for them to manipulate their scores just a few points to score at either side of the threshold. Appendix figure A.3 formally shows there is no evidence of score manipulation at the cutoff point using the McCrary test. The lack of bunching at the threshold also rules out the possibility that teachers nudge students who classified as remedial in second grade to miss the eighth grade ECE to improve the staff's chances of receiving the monetary bonus "Bono Escuela".

Lacking nationally unique IDs in the dataset, Ministry of Education staff matched student scores across grades based on full names (first name, middle name, father's last name and mother's last name). In this setting, 55% of the students who took

the eighth grade test had their second grade scores matched.[9] There are two main reasons for a student in eighth grade not to have his or her second grade test outcomes in the dataset. First, any misspelling prevented matches from being made. Despite this issue, a stricter matching protocol was favored to prevent false matches that could introduce measurement error. Second, some students may have not taken the second grade test because of absence on the test date or because their school did not have enough students in second grade to be tested. While 92% of eighth graders in the country effectively took the test in 2015, the figure for second graders in 2009 was lower, at 81% (Ministerio de Educación, 2009, 2015). The reason for this difference is that the test is administered in all schools that have at least five students enrolled in the respective grade, and some primary schools fail this criterion while most secondary schools meet it. The internal validity of the results of this paper could be threatened by a discontinuous change in the probability of matching at the second grade remedial threshold. It is not possible to test directly for this discontinuity with the data provided by the Ministry, since it includes information for matched students only, so the second grade score for unmatched students is not observable. However, a discontinuity of this sort is highly unlikely, given that students are balanced in the available observable characteristics (Table A.2) together with the lack of bunching at the threshold (Figure A.3). In addition, if the probability of matching changed discontinuously at the second grade math threshold, one would expect this discontinuity to be present also in eighth grade scores, especially given the persistence of the type of changes uncovered in this paper. Table A.3 shows that there is no such discontinuity, further alleviating concerns about internal validity.

6 Conclusions

This paper shows that labels attached to students at a young age can affect learning outcomes persistently, and that this is largely due to inattention to test scores.

[9] The matching rates are 36% for rural males, 40% for rural females, 56% for urban males, and 60% for urban females.

Marginally earning a "remedial" label in a low-stakes second grade math test reduced scores six years later by 0.18 standard deviations (0.11 SD in math, 0.21 SD in language), and substantially increases the likelihood of obtaining "warning" labels in those subjects. These effects are large, comparable in magnitude to the average effects of structured pedagogy, which -in terms of learning- are the most effective educational interventions found in the systematic review by Snilstveit et al. (2015), and are higher than the average effects found for computer interventions and community-based monitoring (Snilstveit et al., 2015).[10] These effects are not only large, they are valid for a considerable proportion of the sample. Given the distribution of scores around the threshold, the optimal bandwidth implies that these results are valid for 26% of rural males in the study sample.

The only previous study on inattention to test scores is Papay, Murnane, and Willett (2016). The authors document the existence of limited attention to scores, but their data do not allow to investigate the mechanisms. Besides being the first study of behavioral inattention to test scores in a developing country, this study is the first to provide evidence on the main mechanisms suggested by the education literature, and shows that classroom environment and household resource reallocation are important driving forces behind the effects on learning outcomes. This suggests a role for parental, peer and teacher inattention that may reinforce the student's own cognitive biases.

This paper has two main limitations worth considering. First, there is a limitation regarding the timing of measurement of the mechanisms. The variables used to investigate outcomes and mechanisms are measured in eighth grade. The lack of information about what happened immediately after the second grade scores and labels were reported to parents and students complicates somewhat the interpretation of the causal chain. For instance, it is not clear if child labor was triggered directly by the second grade label, or if the label triggered a more subtle resource reallocation that

[10] In a systematic review of education interventions in low and middle-income countries, Snilstveit et al. (2015) report that, on average, school feeding programs improve learning in math by 0.10 SD and 0.09 SD in language. Structured pedagogy interventions improve math by 0.14 SD and language by 0.23 SD. Computer interventions improve learning in math by 0.07 SD and language by 0.01 SD. Community based monitoring improve math 0.09 SD and language 0.12 SD.

further reduced the child's school performance, lowering further the expected returns to schooling, and triggering child labor. However, it is possible to conclude that by eighth grade, perception of a negative classroom environment, child labor, and reduced book ownership are forces behind the reduction in test scores described above, and that the changes in these variables have been triggered either directly or indirectly, by marginally obtaining a "remedial" label in second grade math. This calls for further research to pin down the causal chain of the mechanisms. Second, and a consequence of the first one, is that the study cannot provide estimates of inattention parameters, which are theoretically interesting and empirically scarce. This is because the effects on outcomes are result of the combined inattention of parents, teachers, students and their peers, and it is not possible to disentangle the contribution on each of them.

Two main policy implications arise from this study. First is the need for better communication with teachers, parents and students about their scores and how to interpret them. Besides improved communication, schools should provide all stakeholders with tools to improve learning among struggling students. Second, there should be efforts to improve classroom environment, which has been shown to boost learning outcomes (Lavy and Schlosser, 2011; Gutierrez, Molina, and Ñopo, 2018).

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. 2017. "When should you adjust standard errors for clustering?" Tech. rep., National Bureau of Economic Research.

Abaluck, Jason and Abi Adams. 2017. "What do consumers consider before they choose? Identification from asymmetric demand responses." Tech. rep., National Bureau of Economic Research.

Alesina, Alberto, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. 2018. "Revealing Stereotypes: Evidence from immigrants in schools." Tech. rep., National Bureau of Economic Research.

Altonji, Joseph G and Rebecca M Blank. 1999. "Race and gender in the labor market." *Handbook of Labor Economics* 3:3143–3259.

Andersen, Steffen, John Y Campbell, Kasper Meisner Nielsen, and Tarun Ramadorai. 2015. "Inattention and inertia in household finance: Evidence from the Danish mortgage market." Tech. rep., National Bureau of Economic Research.

Anderson, Michael L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American statistical Association* 103 (484):1481–1495.

Avery, Christopher, Oded Gurantz, Michael Hurwitz, and Jonathan Smith. 2018. "Shifting college majors in response to advanced placement exam scores." *Journal of Human Resources* 53 (4):918–956.

Baland, Jean-Marie and James A Robinson. 2000. "Is child labor inefficient?" *Journal of Political Economy* 108 (4):663–679.

Becker, Howard. 1963. *Outsiders: Studies in the Sociology of Deviance*. New York: Free Press.

Brown, Jennifer, Tanjim Hossain, and John Morgan. 2010. "Shrouded attributes and information suppression: Evidence from the field." *The Quarterly Journal of Economics* 125 (2):859–876.

Brown, Sarah and Karl Taylor. 2008. "Bullying, education and earnings: evidence from the National Child Development Study." *Economics of Education Review* 27 (4):387–401.

Busse, Meghan R, Christopher R Knittel, and Florian Zettelmeyer. 2013. "Are consumers myopic? Evidence from new and used car purchases." *American Economic Review* 103 (1):220–56.

Calonico, Sebastian, Matias D Cattaneo, and Max H Farrell. 2018. "On the effect of bias estimation on coverage accuracy in nonparametric inference." *Journal of the American Statistical Association* 113 (522):767–779.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82 (6):2295–2326.

Caplin, Andrew. 2016. "Measuring and modeling attention." *Annual Review of Economics* 8:379–403.

Card, David and A Abigail Payne. 2017. "High school choices and the gender gap in STEM." Tech. rep., National Bureau of Economic Research.

Carlana, Michela. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias." *The Quarterly Journal of Economics* 134 (3):1163–1224. URL `https://doi.org/10.1093/qje/qjz008`.

Cattaneo, Matias D, Rocio Titiunik, and Gonzalo Vazquez-Bare. 2019. "Power calculations for regression-discontinuity designs." *The Stata Journal* 19 (1):210–245.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and taxation: Theory and evidence." *American Economic Review* 99 (4):1145–77.

Cooper, Harris M and Thomas L Good. 1983. *Pygmalion grows up: Studies in the expectation communication process*. Longman Publishing Group.

DellaVigna, Stefano. 2009. "Psychology and economics: Evidence from the field." *Journal of Economic Literature* 47 (2):315–72.

Dizon-Ross, Rebecca. 2019. "Parents' Beliefs About Their Children's Academic Ability: Implications for Educational Investments." *American Economic Review* 109 (8):2728–2765.

Ellison, Glenn. 2005. "A model of add-on pricing." *The Quarterly Journal of Economics* 120 (2):585–637.

Ellison, Glenn and Sara Fisher Ellison. 2009. "Search, obfuscation, and price elasticities on the internet." *Econometrica* 77 (2):427–452.

Eriksen, Tine Louise Mundbjerg, Helena Skyt Nielsen, and Marianne Simonsen. 2014. "Bullying in elementary school." *Journal of Human Resources* 49 (4):839–871.

Gabaix, Xavier. 2019. "Behavioral inattention." *Handbook of Behavioral Economics-Foundations and Applications 2* :261.

Gabaix, Xavier and David Laibson. 2006. "Shrouded attributes, consumer myopia, and information suppression in competitive markets." *The Quarterly Journal of Economics* 121 (2):505–540.

Gelman, Andrew and Guido Imbens. 2019. "Why high-order polynomials should not be used in regression discontinuity designs." *Journal of Business & Economic Statistics* 37 (3):447–456.

Glewwe, Paul and Hanan Jacoby. 1994. "Student achievement and schooling choice in low-income countries: Evidence from Ghana." *Journal of Human Resources* 29 (3):843–865.

Gutierrez, Italo A, Oswaldo Molina, and Hugo Ñopo. 2018. "Stand against bullying: An experimental school intervention." .

INEI. 2016a. "Encuesta Nacional sobre Relaciones Sociales - Principales Resultados." Tech. rep., INEI.

———. 2016b. *Perú: Características de los niños, niñas y adolescentes que trabajan.*

Jacoby, Hanan G. 1994. "Borrowing constraints and progress through school: Evidence from Peru." *The Review of Economics and Statistics* :151–160.

Jussim, Lee and Kent D Harber. 2005. "Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies." *Personality and Social Psychology Review* 9 (2):131–155.

Kristoffersen, Jannie Helene Grøne, Morten Visby Krægpøth, Helena Skyt Nielsen, and Marianne Simonsen. 2015. "Disruptive school peers and student outcomes." *Economics of Education Review* 45:1–13.

Lacetera, Nicola, Devin G Pope, and Justin R Sydnor. 2012. "Heuristic thinking and limited attention in the car market." *American Economic Review* 102 (5):2206–36.

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

Lee, Kenneth, Edward Miguel, and Catherine Wolfram. 2019. "Experimental Evidence on the Economics of Rural Electrification." *Journal of Political Economy* URL `http://dx.doi.org/10.1086/705417`.

Machin, Stephen J, Sandra McNally, and Jenifer Ruiz-Valenzuela. 2018. "Entry Through the Narrow Door: The Costs of Just Failing High Stakes Exams." .

Meza, Fernando and Katherine Pérez. 2018. "El efecto del trabajo infantil sobre los ingresos en la adultez en el sector agropecuario: un análisis para el caso peruano."

Ministerio de Educación. 2009. "Resultados de la Evaluación Censal de Estudiantes 2009." Tech. rep., Ministry of Education.

———. 2015. "Reporte Técnico de la Evaluación Censal de Estudiantes 2015." Tech. rep., Ministry of Education.

Myers, Erica. 2018. "Are Home Buyers Inattentive? Evidence From Capitalization of Energy Costs." *American Economic Journal: Economic Policy: Forthcoming* .

OECD. 2018. "Pisa: Results in focus." Tech. rep., Organisation for Economic Co-operation and Development: OECD.

Papageorge, Nicholas W, Seth Gershenson, and Kyung Min Kang. 2016. "Teacher expectations matter." *Review of Economics and Statistics* :1–46.

Papay, John P, Richard J Murnane, and John B Willett. 2010. "The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts." *Educational Evaluation and Policy Analysis* 32 (1):5–23.

———. 2014. "High-school exit examinations and the schooling decisions of teenagers: Evidence from regression-discontinuity approaches." *Journal of Research on Educational Effectiveness* 7 (1):1–27.

———. 2016. "The impact of test score labels on human-capital investment decisions." *Journal of Human Resources* 51 (2):357–388.

Patrinos, Harry Anthony and George Psacharopoulos. 1997. "Family size, schooling and child labor in Peru–An empirical analysis." *Journal of Population Economics* 10 (4):387–405.

Ravallion, Martin and Quentin Wodon. 2000. "Does child labour displace schooling? Evidence on behavioural responses to an enrollment subsidy." *The Economic Journal* 110 (462):158–175.

Reardon, Sean F, Nicole Arshan, Allison Atteberry, and Michal Kurlaender. 2010. "Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation." *Educational Evaluation and Policy Analysis* 32 (4):498–520.

Rist, Ray. 1977. *On Understanding the Processes of Schooling: The Contributions of Labelling Theory*, vol. Power and Ideology in Education. New York: Oxford University Press, 292–305.

Rosenthal, Robert and Lenore Jacobson. 1968. "Pygmalion in the classroom." *The Urban Review* 3 (1):16–20.

Sempé, Lucas, Patricia Andrade, Flavio Calmet, Beatriz Castillo, Andrés Figallo, Elizabeth Morán, Lucy Muñoz, Fabiola Salvatierra, Osmar Verona et al. 2017. *Evaluación de uso de informes de la Evaluación Censal de Estudiantes en la Escuela*. GRADE - FORGE. URL `http://repositorio.minedu.gob.pe/handle/MINEDU/5444`.

Smith, Jonathan, Michael Hurwitz, and Christopher Avery. 2017. "Giving college credit where it is due: Advanced Placement exam scores and college outcomes." *Journal of Labor Economics* 35 (1):67–147.

Snilstveit, Birte, Jennifer Stevenson, Daniel Phillips, Martina Vojtkova, Emma Gallagher, Tanja Schmidt, Hannah Jobse, Maisie Geelen, Maria Grazia Pastorello, and John Eyers. 2015. "Interventions for improving learning outcomes and access to education in low-and middle-income countries: a systematic review." Tech. rep., International Initiative for Impact Evaluation.

Spencer, Steven J, Claude M Steele, and Diane M Quinn. 1999. "Stereotype threat and women's math performance." *Journal of Experimental Social Psychology* 35 (1):4–28.

Taubinsky, Dmitry and Alex Rees-Jones. 2017. "Attention variation and welfare: theory and evidence from a tax salience experiment." *The Review of Economic Studies* 85 (4):2462–2496.

Table 1: Descriptive statistics

Variable	Males, Rural	Females, Rural	Males, Urban	Females, Urban
Number of Observations	10,543	10,185	123,436	129,188
% of Subsample	0.3585	0.4037	0.5584	0.6045
Indigenous language	0.2381	0.2429	0.0393	0.0400
	(0.4259)	(0.4288)	(0.1944)	(0.1960)
Mother has secondary education	0.1784	0.1641	0.6471	0.6072
	(0.3829)	(0.3704)	(0.4779)	(0.4884)
Dads education secondary	0.3040	0.3061	0.7290	0.7172
	(0.4600)	(0.4609)	(0.4445)	(0.4504)
Socio Economic Index	-1.1294	-1.1600	0.3376	0.2566
	(0.7991)	(0.7645)	(0.8524)	(0.8618)
Remedial math, 2nd grade	0.5456	0.5590	0.3143	0.3470
	(0.4979)	(0.4965)	(0.4642)	(0.4760)
In transition math, 2nd grade	0.3577	0.3451	0.4526	0.4665
	(0.4794)	(0.4754)	(0.4977)	(0.4989)
Proficient math, 2nd grade	0.0965	0.0954	0.2327	0.1861
	(0.2953)	(0.2938)	(0.4226)	(0.3892)
Remedial language, 2nd grade	0.3440	0.3274	0.1091	0.0957
	(0.4751)	(0.4693)	(0.3118)	(0.2942)
In transition language, 2nd grade	0.5373	0.5392	0.5756	0.5494
	(0.4986)	(0.4985)	(0.4943)	(0.4976)
Proficient language, 2nd grade	0.1187	0.1334	0.3152	0.3549
	(0.3234)	(0.3400)	(0.4646)	(0.4785)
Warning math, 8th grade	0.5427	0.6174	0.2484	0.2947
	(0.4982)	(0.4861)	(0.4321)	(0.4559)
Remedial math, 8th grade	0.3654	0.3109	0.4170	0.4372
	(0.4816)	(0.4629)	(0.4931)	(0.4960)
In transition math, 8th grade	0.0617	0.0504	0.1717	0.1546
	(0.2407)	(0.2187)	(0.3771)	(0.3615)
Proficient math, 8th grade	0.0302	0.0213	0.1629	0.1135
	(0.1710)	(0.1444)	(0.3693)	(0.3172)
Warning language, 8th grade	0.4793	0.4854	0.1401	0.1361
	(0.4996)	(0.4998)	(0.3471)	(0.3429)
Remedial language, 8th grade	0.3987	0.3897	0.3799	0.3737
	(0.4896)	(0.4877)	(0.4854)	(0.4838)
In transition language, 8th grade	0.0957	0.0941	0.2787	0.2763
	(0.2942)	(0.2919)	(0.4484)	(0.4472)
Proficient language, 8th grade	0.0264	0.0308	0.2012	0.2139
	(0.1602)	(0.1729)	(0.4009)	(0.4101)

Notes: Standard deviations in parentheses. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table 2: Descriptive statistics: mechanisms suggested by the literature

Variable	Males, Rural	Females, Rural	Males, Urban	Females, Urban
Interaction with family index	-0.0500	0.0252	-0.0184	0.0688
	(0.9947)	(1.0018)	(0.9983)	(1.0033)
Perception of own skills index	-0.0064	0.0921	-0.1491	-0.0041
	(0.9985)	(1.0180)	(0.9515)	(0.9990)
Negative classroom environment	0.1536	0.0138	-0.0121	-0.1159
	(1.1213)	(1.0397)	(0.9872)	(0.8958)
Five or more books at home	0.7269	0.7629	0.8881	0.9012
	(0.4456)	(0.4253)	(0.3152)	(0.2983)
Would dropout if possible	0.4114	0.3843	0.2083	0.1730
	(0.4921)	(0.4864)	(0.4061)	(0.3782)
Would dropout because of work	0.4106	0.3141	0.1815	0.1104
	(0.4920)	(0.4642)	(0.3855)	(0.3134)

Notes: Standard deviations in parentheses. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table 3: Effect of being classified as remedial in second grade math on eighth grade performance, rural males

Subject	Standardized score	% Warning	% Remedial	% In transition
Overall	-0.175	0.144	0.009	0.021
	(0.067)	(0.043)	(0.029)	(0.018)
Math	-0.109	0.099	-0.001	0.013
	(0.058)	(0.056)	(0.025)	(0.013)
Language	-0.208	0.180	-0.012	0.011
	(0.074)	(0.045)	(0.032)	(0.011)
Total Score				
Mean	-0.625	0.386	0.860	0.967
Bias-robust p-value	0.011	0.002	0.847	0.304
FDR q-value	0.017	0.009	0.863	0.527
Math				
Mean	-0.460	0.562	0.924	0.979
Bias-robust p-value	0.077	0.097	0.863	0.345
FDR q-value	0.077	0.291	0.863	0.527
Language				
Mean	-0.699	0.494	0.896	0.982
Bias-robust p-value	0.005	0.000	0.651	0.351
FDR q-value	0.015	0.001	0.837	0.527
Observations	9,523	9,523	9,523	9,523

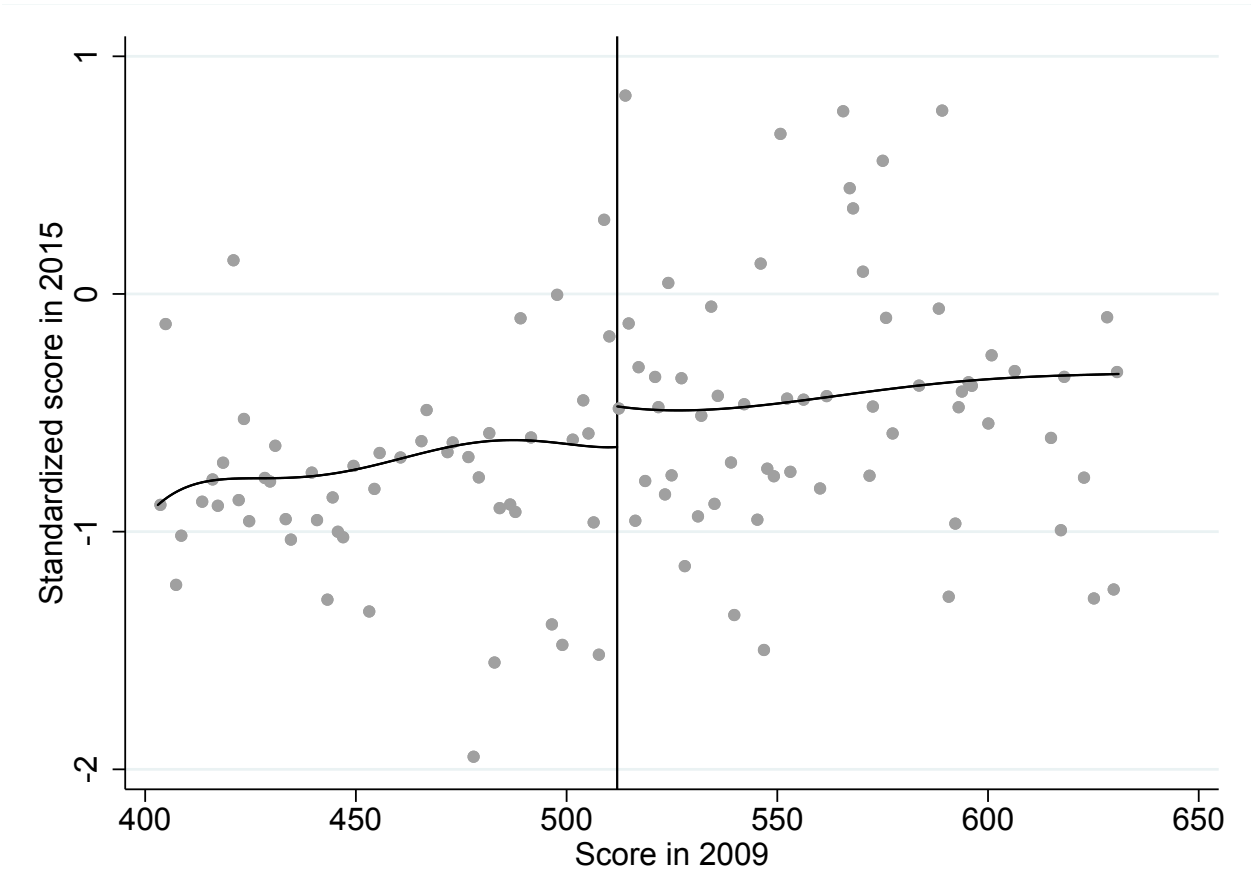
Notes: The dependent variable in column 1 is the standardized score in eighth grade, in columns 2-4 is the probability of scoring at or below the level indicated by the columns. The row indicates the subject. The first row indicates the effect on total score (column 1) and the probability of scoring at or below the level indicated by the column in both subjects (columns 2-4). Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table 4: Mechanisms, rural males

	Family Support	Academic Skills	Negative class environment	work	books at home
Math score < 512	-0.116	-0.154	0.291	0.218	-0.462
	(0.381)	(0.132)	(0.127)	(0.081)	(0.187)
Mean	-0.056	0.012	0.168	0.416	0.725
Bias-robust p-value	0.805	0.265	0.019	0.004	0.015
FDR q-value	0.805	0.332	0.032	0.002	0.032
Observations	8,419	7,887	7,393	4,795	9,006

Notes: The outcome variable is indicated by the column titles. The first three columns are indices of family support, perception about own academic skills and perception of a negative classroom environment. Column (4) is an indicator of child labor. Column (5) is an indicator of the student's household owning five or more books. Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

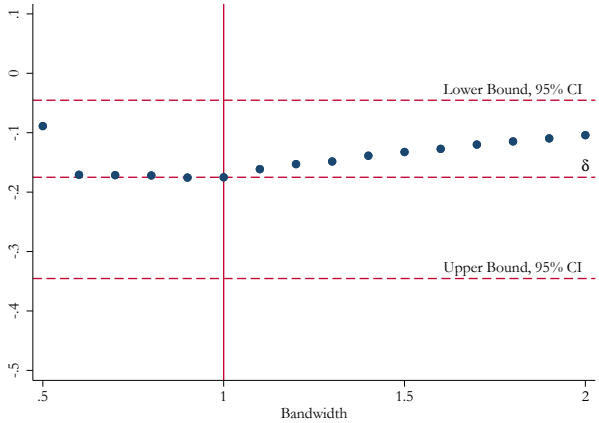
Figure 1: Second grade math scores and eighth grade score



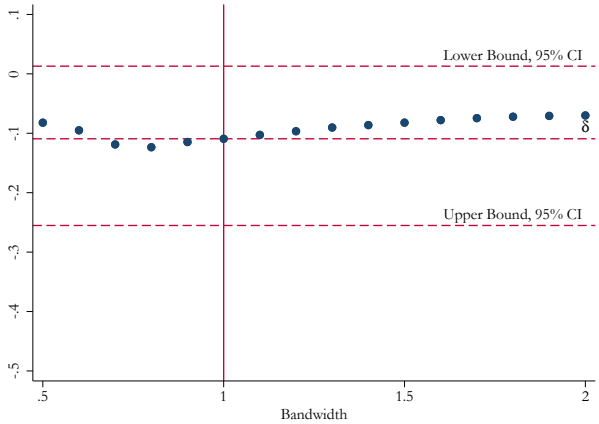
Discontinuity at the threshold = -0.18 SD. FDR q-value = 0.04.

Figure 2: Robustness to bandwidth selection - main outcomes

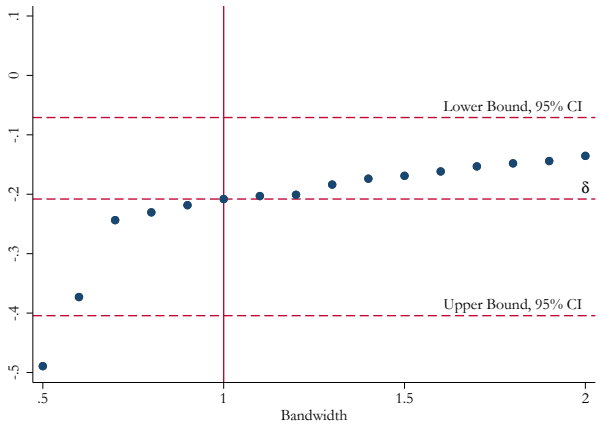
(a) Overall Score



(b) Math



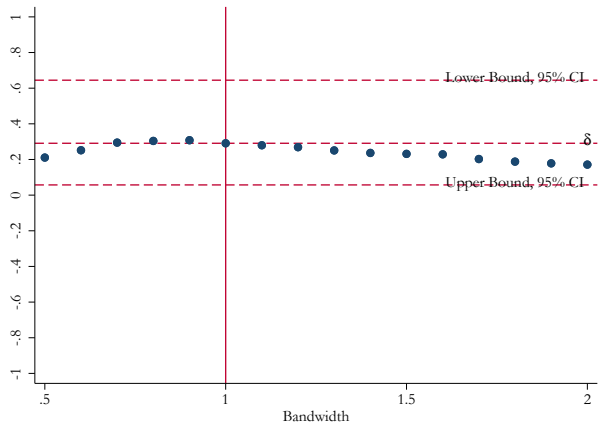
(c) Language



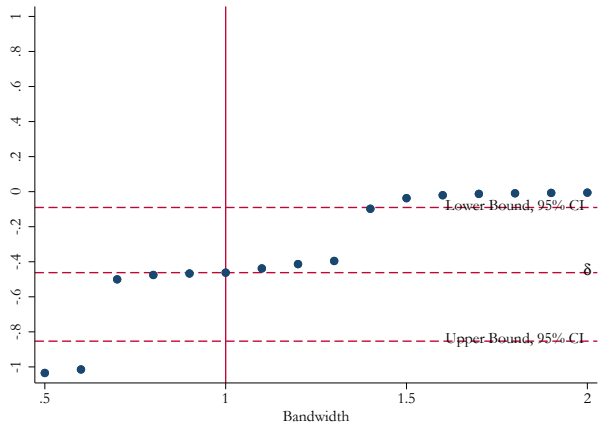
Notes: The vertical axis measures the estimated RD effect (in standard deviations) with the bandwidth indicated in the horizontal axis. Bandwidths range from 50% to 200% of the optimal bandwidth selected by (Calonico, Cattaneo, and Titiunik, 2014; Calonico, Cattaneo, and Farrell, 2018). The solid vertical line indicates the optimal bandwidth. Dotted lines indicate the estimated RD effect and its confidence interval resulting from the optimal bandwidth. Source: 2009 and 2015 *ECE*, Ministry of Education

Figure 3: Robustness to bandwidth selection - mechanisms

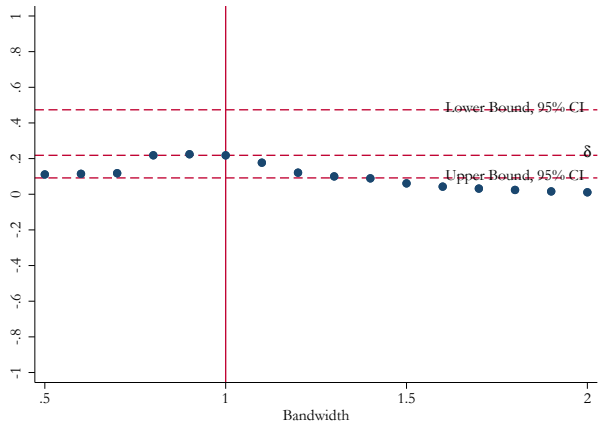
(a) Negative classroom environment



(b) Book ownership

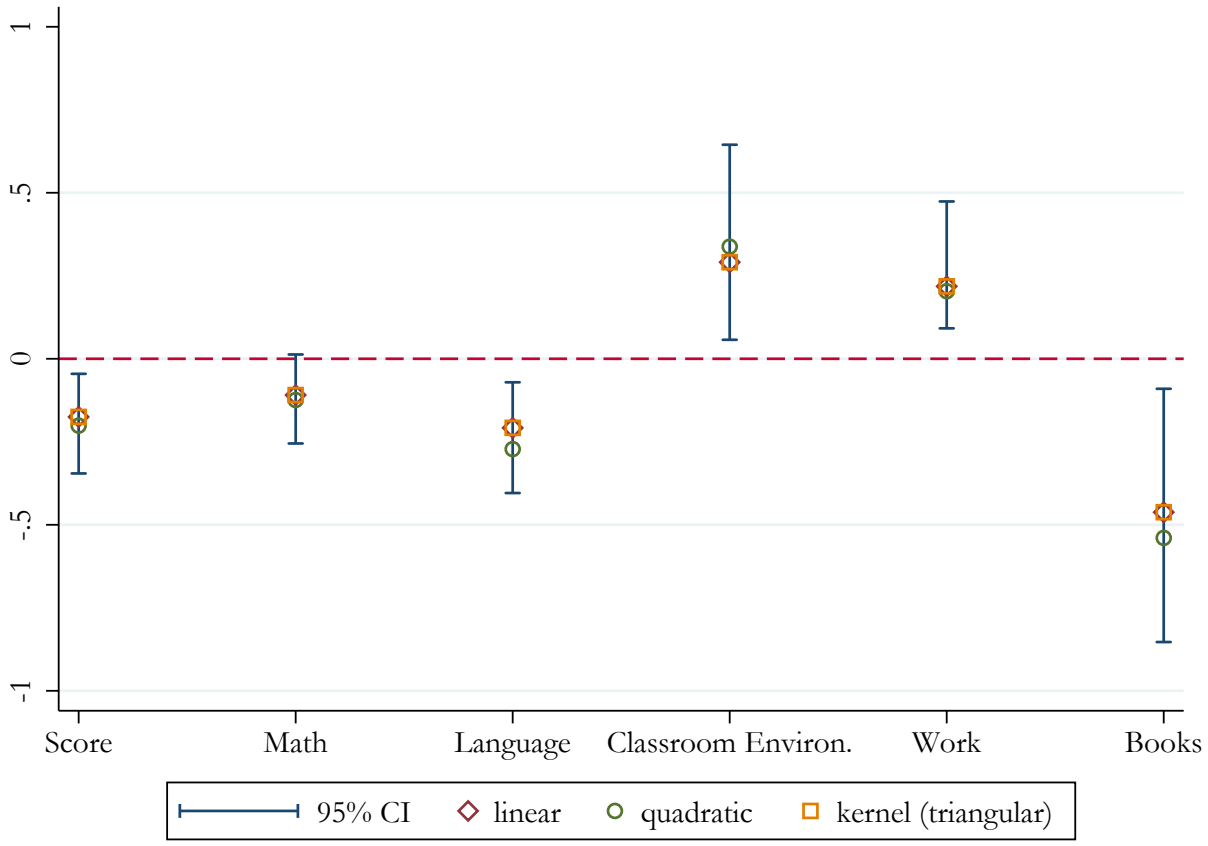


(c) Child labor



Notes: The vertical axis measures the estimated RD effect (in standard deviations) with the bandwidth indicated in the horizontal axis. Bandwidths range from 50% to 200% of the optimal bandwidth selected by (Calonico, Cattaneo, and Titiunik, 2014; Calonico, Cattaneo, and Farrell, 2018). The solid vertical line indicates the optimal bandwidth. Dotted lines indicate the estimated RD effect and its confidence interval resulting from the optimal bandwidth. Source: 2009 and 2015 *ECE*, Ministry of Education

Figure 4: Robustness to polynomial degree



Notes: The vertical axis indicates the estimated RD effect resulting from estimating equation (1) with local polynomials of degrees one, two, and with the triangular kernel. The bias-corrected confidence intervals are reported for the local polynomial of degree one. Source: 2009 and 2015 *ECE*, Ministry of Education

Online Appendix

Table A.1: Descriptive statistics (unmatched sample)

Variable	Males, Rural	Females, Rural	Males, Urban	Females, Urban
Number of Observations	18,867	15,046	97,625	84,517
% of Subsample	0.6415	0.5963	0.4416	0.3955
Indigenous language	0.3638	0.4009	0.0979	0.1039
	(0.4811)	(0.4901)	(0.2972)	(0.3051)
Moms education secondary	0.1452	0.1333	0.4975	0.4695
	(0.3523)	(0.3399)	(0.5000)	(0.4991)
Dads education secondary	0.2575	0.2621	0.6050	0.6035
	(0.4372)	(0.4398)	(0.4889)	(0.4892)
Socio Economic Index	-1.2637	-1.3161	-0.0058	-0.0785
	(0.7728)	(0.7361)	(0.9520)	(0.9544)
Warning math, 8th grade	0.6511	0.7336	0.3958	0.4616
	(0.4766)	(0.4421)	(0.4890)	(0.4985)
Remedial math, 8th grade	0.2891	0.2215	0.4164	0.3923
	(0.4534)	(0.4153)	(0.4930)	(0.4883)
In transition math, 8th grade	0.0412	0.0312	0.1153	0.0941
	(0.1987)	(0.1738)	(0.3194)	(0.2920)
Proficient math, 8th grade	0.0186	0.0137	0.0725	0.0519
	(0.1351)	(0.1162)	(0.2592)	(0.2218)
Warning comm, 8th grade	0.5954	0.6342	0.2611	0.2686
	(0.4908)	(0.4817)	(0.4393)	(0.4432)
Remedial comm, 8th grade	0.3297	0.2974	0.4257	0.4117
	(0.4701)	(0.4571)	(0.4945)	(0.4921)
In transition comm, 8th grade	0.0594	0.0540	0.2088	0.2026
	(0.2364)	(0.2260)	(0.4065)	(0.4020)
Proficient comm, 8th grade	0.0155	0.0144	0.1043	0.1171
	(0.1234)	(0.1192)	(0.3056)	(0.3216)

Notes: Standard deviations in parentheses. Source: 2015 *ECE*, Ministry of Education.

Table A.2: Balance at the "remedial" threshold

Variable	Males, Rural	Females, Rural	Males, Urban	Females, Urban
Socioeconomic Index	-0.1846	-0.0543	0.0211	-0.0033
	(0.1080)	(0.0863)	(0.0139)	(0.0163)
Mom indigenous	0.0389	0.0152	-0.0028	0.0016
	(0.0974)	(0.0562)	(0.0038)	(0.0038)
Dad indigenous	0.0724	0.0003	-0.0018	0.0005
	(0.0613)	(0.0498)	(0.0030)	(0.0028)
Mom secondary	-0.0452	-0.0043	0.0156	-0.0022
	(0.0472)	(0.0352)	(0.0083)	(0.0080)
Dad secondary	-0.0414	0.0091	0.0048	-0.0060
	(0.0413)	(0.0663)	(0.0087)	(0.0079)
Attended preschool	-0.0588	0.0546	-0.0015	-0.0004
	(0.0417)	(0.0549)	(0.0058)	(0.0050)
Bias-robust p-value, SE index	0.1316	0.5838	0.6463	0.3845
Bias-robust p-val, mom indig	0.7316	0.8871	0.4431	0.2571
Bias-robust p-val, dad indig	0.3655	0.9874	0.9077	0.3525
Bias-robust p-val, mom secondary	0.3489	0.9783	0.1948	0.3329
Bias-robust p-val, dad secondary	0.3136	0.9190	0.9192	0.1555
Bias-robust p-val, preschool	0.2430	0.3142	0.6274	0.5581
Observations	9,449	9,144	93,273	103,702

Notes: Each cell reports the regression discontinuity estimate at the remedial threshold in second grade math on the variable indicated by the row, for the sample indicated in the column. Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2015 *ECE*, Ministry of Education.

Table A.3: Change in probability of attrition at eighth grade math score threshold

Variable	Males, Rural	Females, Rural	Males, Urban	Females, Urban
Math Score<512	0.0163	0.0060	0.0106	-0.0038
	(0.0179)	(0.0187)	(0.0093)	(0.0070)
Bias-robust p-value	0.3725	0.7040	0.1917	0.4718
Mean	0.6444	0.5986	0.4705	0.4142
Number of observations	28,599	24,716	189,534	191,216

Notes: Each column reports the regression discontinuity estimate at the remedial threshold in eighth grade math on the probability of a student having his or her data matched with the second grade results. Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2015 *ECE*, Ministry of Education.

Table A.4: Effect of being classified as remedial in second grade math on eighth grade performance category, rural females

Subject	Standardized score	% Warning	% Remedial	% In transition
Overall	0.033	-0.040	0.011	0.002
	(0.058)	(0.053)	(0.041)	(0.024)
Math	0.357	-0.249	-0.011	-0.028
	(0.264)	(0.147)	(0.106)	(0.038)
Language skills	0.041	0.002	0.009	0.016
	(0.071)	(0.042)	(0.026)	(0.013)
Mean, overall	-0.685	0.422	0.866	0.967
Bias-robust p-value, overall	0.586	0.458	0.678	0.838
Mean, math	-0.576	0.635	0.942	0.985
Bias-robust p-value, math	0.204	0.105	0.932	0.501
Mean, languange skills	-0.694	0.498	0.889	0.977
Bias-robust p-value, language	0.703	0.835	0.669	0.168
Observations	9,209	9,209	9,209	9,209

Notes: The dependent variable in column 1 is the standardized score in eighth grade, in columns 2-4 is the probability of scoring at or below the level indicated by the columns. The row indicates the subject. The first row indicates the effect on total score (column 1) and the probability of scoring at or below the level indicated by the column in both subjects (columns 2-4). Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table A.5: Effect of being classified as remedial in second grade math on eighth grade performance category, urban males

Subject	Standardized score	% Warning	% Remedial	% In transition
Overall	0.006	-0.001	-0.003	0.001
	(0.013)	(0.007)	(0.009)	(0.006)
Math	-0.003	-0.002	-0.003	-0.005
	(0.015)	(0.008)	(0.008)	(0.005)
Language skills	0.013	0.000	-0.002	0.003
	(0.014)	(0.009)	(0.008)	(0.005)
Mean, overall	0.058	0.122	0.561	0.840
Bias-robust p-value, overall	0.754	0.732	0.528	0.725
Mean, math	0.064	0.305	0.772	0.922
Bias-robust p-value, math	0.806	0.902	0.480	0.238
Mean, languange skills	0.043	0.172	0.612	0.879
Bias-robust p-value, language	0.483	0.495	0.548	0.952
Observations	94,719	94,719	94,719	94,719

Notes: The dependent variable in column 1 is the standardized score in eighth grade, in columns 2-4 is the probability of scoring at or below the level indicated by the columns. The row indicates the subject. The first row indicates the effect on total score (column 1) and the probability of scoring at or below the level indicated by the column in both subjects (columns 2-4). Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table A.6: Effect of being classified as remedial in second grade math on eighth grade performance category, urban females

Subject	Standardized score	% Warning	% Remedial	% In transition
Overall	0.000	-0.003	0.001	0.003
	(0.012)	(0.006)	(0.008)	(0.006)
Math	-0.000	0.004	0.001	-0.006
	(0.012)	(0.009)	(0.006)	(0.004)
Language skills	-0.002	-0.001	0.003	0.005
	(0.014)	(0.007)	(0.008)	(0.005)
Mean, overall	0.049	0.122	0.545	0.828
Bias-robust p-value, overall	0.977	0.734	0.722	0.918
Mean, math	-0.035	0.341	0.809	0.941
Bias-robust p-value, math	0.770	0.554	0.656	0.048
Mean, languange skills	0.127	0.158	0.575	0.848
Bias-robust p-value, language	0.670	0.523	0.756	0.685
Observations	105,135	105,135	105,135	105,135

Notes: The dependent variable in column 1 is the standardized score in eighth grade, in columns 2-4 is the probability of scoring at or below the level indicated by the columns. The row indicates the subject. The first row indicates the effect on total score (column 1) and the probability of scoring at or below the level indicated by the column in both subjects (columns 2-4). Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table A.7: Effects of crossing other thresholds on eighth-grade standardized scores

Threshold	Rural Males	Rural Females	Urban Males	Urban Females
Math, Proficient	-0.0279	0.1137	-0.0269	0.0048
	(0.0758)	(0.0875)	(0.0134)	(0.0129)
Language, Remedial	0.0280	-0.0005	0.0030	0.0350
	(0.0527)	(0.0567)	(0.0171)	(0.0162)
Language, Proficient	0.1319	0.0268	-0.0072	-0.0346
	(0.1004)	(0.0902)	(0.0137)	(0.0122)
Bias-robust p-val, math profic	0.780	0.249	0.298	0.804
Bias-robust p-val, lang remedial	0.513	0.895	0.921	0.048
Bias-robust p-val, lang profic	0.479	0.668	0.621	0.870
Observations	4,791	4,484	84,561	84,272

Notes: Each cell reports the estimated discontinuity in standardized eighth-grade scores at the second-grade threshold indicated by the row title, for the subpopulation indicated by the column. Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table A.8: Mechanisms, rural females

	Family Support	Academic Skills	Negative class environment	work	books at home
Math score < 512	0.041	-0.166	-0.111	0.016	0.050
	(0.097)	(0.092)	(0.140)	(0.079)	(0.053)
Mean	0.024	0.104	0.027	0.316	0.758
Bias-robust p-value	0.764	0.070	0.411	0.879	0.421
Observations	7,914	7,383	6,702	4,482	8,740

Notes: The outcome variable is indicated by the column titles. The first three columns are indices of family support, perception about own academic skills and perception of a negative classroom environment. Column (4) is an indicator of child labor. Column (5) is an indicator of the student's household owning five or more books. Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table A.9: Mechanisms, urban males

	Family Support	Academic Skills	Negative class environment	work	books at home
Math score < 512	0.027	-0.037	-0.021	-0.011	0.009
	(0.018)	(0.021)	(0.019)	(0.016)	(0.006)
Mean	-0.051	-0.064	0.008	0.190	0.875
Bias-robust p-value	0.245	0.054	0.299	0.493	0.194
Observations	86,226	80,986	77,219	29,691	89,829

Notes: The outcome variable is indicated by the column titles. The first three columns are indices of family support, perception about own academic skills and perception of a negative classroom environment. Column (4) is an indicator of child labor. Column (5) is an indicator of the student's household owning five or more books. Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table A.10: Mechanisms, urban females

	Family Support	Academic Skills	Negative class environment	work	books at home
Math score < 512	0.033	-0.010	0.007	0.002	0.011
	(0.017)	(0.017)	(0.017)	(0.013)	(0.005)
Mean	0.042	0.063	-0.107	0.114	0.892
Bias-robust p-value	0.208	0.164	0.638	0.821	0.033
Observations	95,491	90,221	84,431	29,870	100,317

Notes: The outcome variable is indicated by the column titles. The first three columns are indices of family support, perception about own academic skills and perception of a negative classroom environment. Column (4) is an indicator of child labor. Column (5) is an indicator of the student's household owning five or more books. Heteroskedasticity-robust covariance matrix with at least three nearest neighbors. Source: 2009 and 2015 *ECE*, Ministry of Education.

Table A.11: Components of "interaction with family" index

Statement	Never	Sometimes	Frequently	Always
My parents and I talk about homework.	0	0	1	1
My parents help me with homework.	0	0	1	1
My parents explain to me school topics in any school subject.	0	0	1	1
My parents are aware of my grades.	0	0	1	1
My parents recommend me books that I have not read yet.	0	0	1	1
My parents and I talk about how I am doing in school.	0	0	1	1
My parents and I talk about my friends and what we do together.	0	0	1	1
My parents and I talk about things that worry me.	0	0	1	1
My parents and I practice sports together.	0	0	1	1
My parents and I talk about radio and TV shows, books or movies.	0	0	1	1
My parents and I talk about community, country or world affairs.	0	0	1	1
My parents and I attend cultural activities together (festivals, parades, etc.).	0	0	1	1
My parents and I attend community activities.	0	0	1	1
My parents and I go to museums and expos in galeries or in the street.	0	0	1	1
My parents and I go to the movies or theatre.	0	0	1	1

Notes: The first column shows each question in the index. Columns 2-5 show which values are taken as 1 in the transformed variable. See main text for details. Source: 2015 *ECE*, Ministry of Education.

Table A.12: Components of "perception of own academic skills" index

Statement	Never	Sometimes	Frequently	Always
I can understand any math-related topic.	0	0	1	1
When they teach us a new topic in math I can learn it promptly.	0	0	1	1
I think I can understand hard topics in math.	0	0	1	1
When I take a math test, I am sure I will be able to get the right answers.	0	0	1	1
I can help my classmates understand our math homework.	0	0	1	1
I can do math homework without help.	0	0	1	1
I am confident I can pass math without trouble.	0	0	1	1
I am good solving math problems.	0	0	1	1
I feel more capable the more math I learn.	0	0	1	1
I feel I am good at math.	0	0	1	1
I feel confident studying what they teach me in language skills.	0	0	1	1
I can answer correctly if asked about any topic in language skills.	0	0	1	1
I am able to understand everything they teach us in language skills.	0	0	1	1
I can solve individual homework of language skills withough help.	0	0	1	1
I feel confident when I take a language skills test.	0	0	1	1
I can help my classmates with language skills homework they find difficult.	0	0	1	1
When I participate in language skills class I trust I will do fine.	0	0	1	1
I have good performance in language skills.	0	0	1	1
It is easy for me to understand what I read.	0	0	1	1
I am confident I will pass language skills without trouble.	0	0	1	1

Notes: The first column shows each question in the index. Columns 2-5 show which values are taken as 1 in the transformed variable. See main text for details. Source: 2015 *ECE*, Ministry of Education.

Table A.13: Components of "Negative classroom environment" index

Statement	Never	Sometimes	Frequently	Always
My teachers allow my classmates to be noisy during class.	0	0	1	1
My teachers solve class conflicts by talking thtough them.	1	1	0	0
My teachers let students misbehave.	0	0	1	1
My teachers explain to us the reasons why we should behave.	1	1	0	0
My teachers yell at us when we misbehave.	0	0	1	1
My teachers encourage us to study.	1	1	0	0
My teachers give us additional help when we need it.	1	1	0	0
My teachers congratulate us when we do a good job.	1	1	0	0
My teachers care for our wellbeing.	1	1	0	0
My teachers make me feel bad when I make a mistake.	0	0	1	1
My teachers listen carefully my opinions.	1	1	0	0
My teachers explain gently when I don't understand something.	1	1	0	0
My teachers respect my opinions.	1	1	0	0
There is a bad relation between students and teachers.	0	0	1	1
I have a good relation with my teachers.	1	1	0	0
My teachers critize me in front of my classmates when I make mistakes.	0	0	1	1
In school we are encouraged to be friendly with our peers.	1	1	0	0
I feel uncomfortable and out of place in school.	0	0	1	1
At school we are encouraged to respect our peers.	1	1	0	0
When a student bullies someone, nobody speaks up.	0	0	1	1

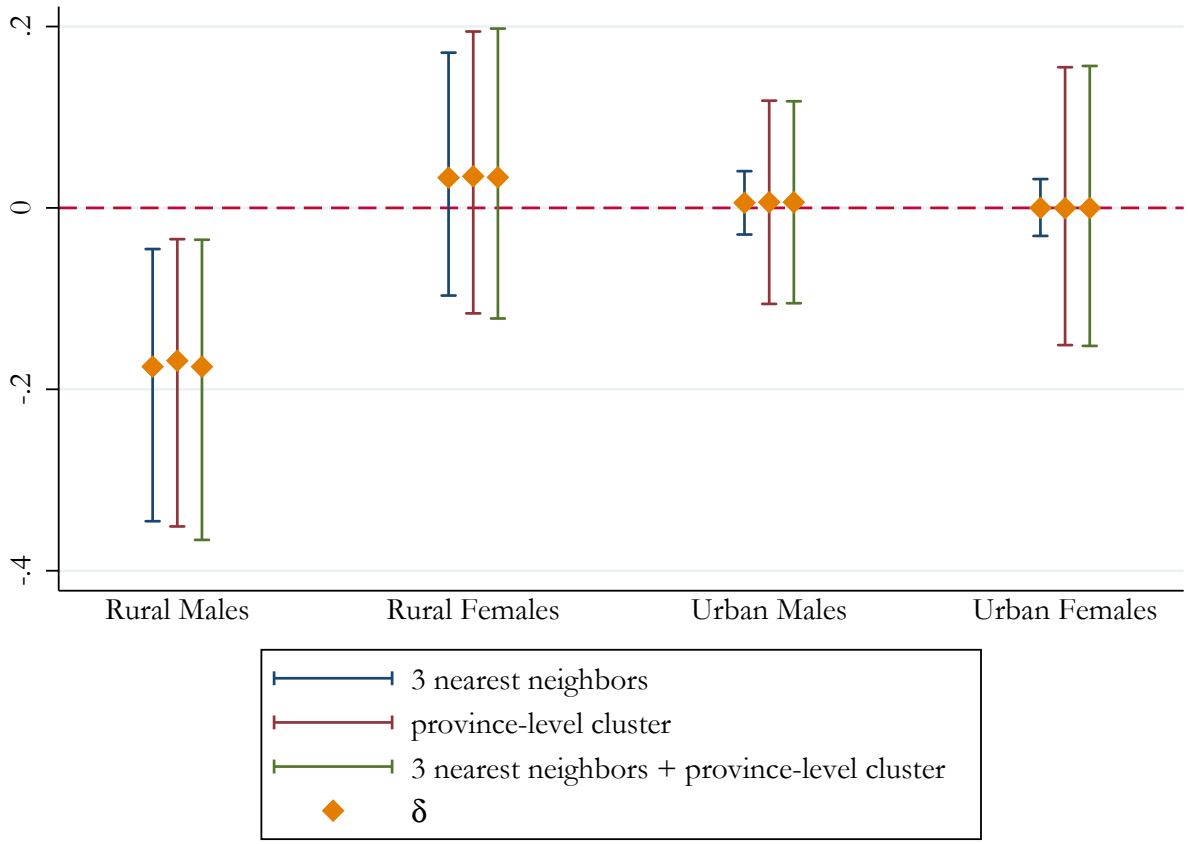
Notes: The first column shows each question in the index. Columns 2-5 show which values are taken as 1 in the transformed variable. See main text for details. Source: 2015 *ECE*, Ministry of Education.

Table A.14: Ex-post power calculations

			Rural		Urban	
Subject	Threshold	Effect	Males	Females	Males	Females
Math	Remedial	MDE	0.164	0.160	0.114	0.156
		$\hat{\delta}$	0.175	0.033	0.006	0.001
	Proficient	MDE	0.205	0.272	0.205	0.244
		$\hat{\delta}$	0.028	0.114	0.027	0.005
Language	Remedial	MDE	0.154	0.153	0.096	0.097
		$\hat{\delta}$	0.028	0.001	0.003	0.035
	Proficient	MDE	0.401	0.441	0.192	0.177
		$\hat{\delta}$	0.132	0.027	0.007	0.035

Notes: This table reports minimum detectable effects (MDE) from ex-post power calculations with 80% power and 95% confidence (Cattaneo, Titiunik, and Vazquez-Bare, 2019). The estimated discontinuities (in absolute value) are reported for reference. Source: 2015 *ECE*, Ministry of Education.

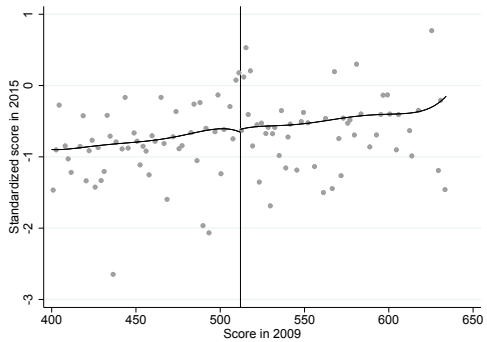
Figure A.1: Effect of being classified as remedial in second grade math on eighth grade scores



Notes: The figure presents estimates for δ_1 in equation (1). The outcome variable is the standardized test score in 8th grade. Bias-robust confidence intervals are calculated with three methods: heteroskedasticity-robust covariance matrix with at least three nearest neighbors, clustered standard errors at the province level, and the combination of both. Source: 2009 and 2015 *ECE*, Ministry of Education.

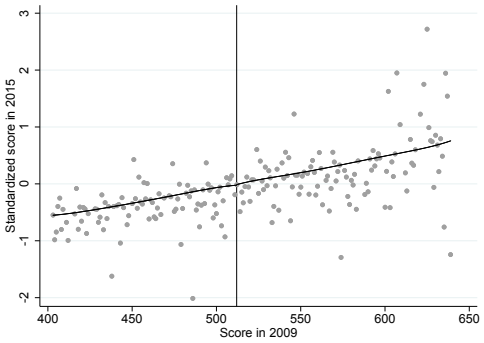
Figure A.2: Second grade math scores and eighth grade score

(a) Rural Females



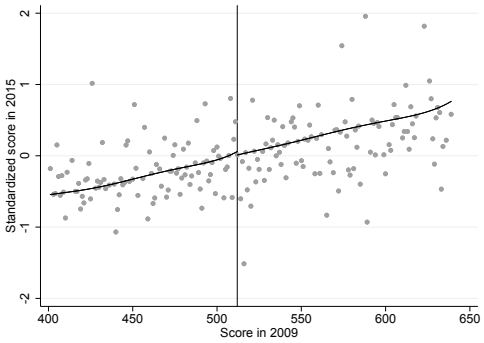
Discontinuity at the threshold = -0.03 SD.
FDR q-value = 0.98.

(b) Urban Males



Discontinuity at the threshold = 0.01 SD.
FDR q-value = 0.98.

(c) Urban Females



Discontinuity at the threshold = -0.01 SD.
FDR q-value = 0.98.

Notes: The horizontal axis shows the score in second grade math. The vertical axis measures the standardized overall score in eighth grade. Source: 2009 and 2015 *ECE*, Ministry of Education

Figure A.3: McCrary Test, Rural Males