Non-Cognitive Skills and Remedial Education: Good News for Girls*

Marianna Battaglia
University of Alicante†

Marisa Hidalgo-Hidalgo
University Pablo de Olavide‡

September 1, 2020

Abstract

Growing evidence shows that non-cognitive skills are crucial for labor market and other outcomes in life. However, little is known about the role of education in improving these abilities, especially for disadvantaged teenagers in developed countries. We address two questions: can remedial educational interventions improve their non-cognitive skills? and, can we expect heterogeneous effects by gender? We take advantage of a remedial program for under-performing students implemented in Spain and we consider testing behaviors as measures of non-cognitive skills. The impact of remedial programs on these abilities, especially malleable for teenagers, has been overlooked in the literature. We find that the program had a substantial positive effect on girls' non-cognitive skills but not on boys'.

Keywords: remedial education, test performance, PISA
JEL classification codes: H52, I23, I28, J24

*We are grateful to Pau Balart, Cristina Borra, Kevin Denny, Paul Devereux, Sue Dynarski, Chris Jepsen and Anna Sanz-de-Galdeano, for helpful suggestions. We benefited from valuable comments of seminar participants at 2018 UCD Workshop on the Economics of Education, Alicante Workshop on Quantitative Economics, IZA World Labor Conference 2018, Seville Workshop on Remedial Education, Non-cognitive skills and Gender, 2018 ASSET Meeting, ESPE 2019, 4th DONDENA Workshop on Public Policy 2019, AGEW 2020 and Australian National University. We gratefully acknowledge Ismael Sanz and F. Javier García Crespo (INEE) for their help with the data. Financial support from Fundación Ramón Areces and Ministerio de Economía y Competitividad (ECO2017-83069-P) is gratefully acknowledged. All opinions expressed are of the authors, all errors are our own.

†Corresponding author. Department of Economics (FAE), University of Alicante, Campus de San Vicente, 03080 Alicante, Spain. Tel: +34 965 90 3400 (ext: 3218). Contact email: mbattaglia@ua.es

‡Department of Economics, University Pablo de Olavide, Ctra. Utrera, Km.1, 41013, Seville, Spain. Tel: +34 954 977 982. Contact email: mhidalgo@upo.es

1 Introduction

Remedial education programs are designed to help poor-performing students to satisfy minimum academic standards. This is usually achieved by means of a targeted increase in instruction time combined with after-school individualized teaching in small study groups. These types of interventions are currently subject to increasing interest, especially in developed countries, where reducing NEET (not in employment, education or training) rates has become a great challenge.[1] Still, policies targeting low-performing students are generally difficult to evaluate due to sample selection, as children with learning difficulties are not randomly assigned to programs. Only a few works address the identification problem and usually document the effectiveness of these programs in improving cognitive outcomes (see Lavy and Schlosser (2005) or De Paola and Scoppa (2015) among others). However, the effect of remedial education programs on non-cognitive abilities and its possible heterogeneous effects have so far been rarely investigated. This is precisely the aim of this paper.

In skill acquisition both cognitive and non-cognitive abilities are relevant in explaining long-term outcomes such as higher education investment and job market prospects. A growing body of the literature suggests that non-cognitive skills are as crucial as cognitive skills in determining students' school achievements and in turn their educational choices (Heckman and Rubinstein, 2001; Heckman et al., 2006; Cunha and Heckman, 2008; Lindqvist and Westman, 2011). Moreover, as suggested by Carneiro and Heckman (2003) and Almlund et al. (2011), both cognitive and non-cognitive skills differ in their malleability over the life cycle, with the latter being more malleable than the former ones at later ages. Abilities other than cognitive can therefore be relevant when teenagers are involved in policy interventions such as remedial education programs, with lasting consequences in the long-term. Interestingly, and similarly important for our study, recent literature documents both the existence of a positive gender gap favoring girls in several measures of cognitive and non-cognitive skills and that non-cognitive returns to exter-

[1] See Heckman (2000) for a review on interventions during the nineties in the US and Carcillo et al. (2015) for a summary of interventions in other OECD countries.

nal inputs differ markedly by gender.[2] In this paper, we take into account the previous empirical facts and address the following questions: (i) Can educational interventions improve non-cognitive skills? (ii) Can we expect heterogeneous effects depending on the students' gender?

We provide new evidence on these questions by taking advantage of two events. First, a program that offered remedial education for under-performing students from poor socioeconomic backgrounds, the Program for School Guidance (PAE), which was implemented in Spain between 2005 and 2012.[3] It explicitly focused on studying habits and organization techniques. And second, the availability of non-self-assessed measures of students' non-cognitive skills for a representative sample of Spanish adolescents in 2012. Similar to recent literature (see Balart et al. (2018), among others, and the review below), we use the term non-cognitive skills to describe the personal attributes not thought to be measured by IQ or standardized tests. They are "patterns of thought, feelings, and behavior" (Borghans et al., 2008), such as academic perseverance and learning strategies, which differ from the ability to "perform higher mental processes of reasoning, remembering, understanding, and problem solving" (Bernstein et al., 2007). We thus consider testing and survey behaviors, for instance decline in performance during the test, as measures of non-cognitive skills. Data on these measures are obtained from external evaluations of the schools, the Programme for International Student Assessment (PISA) 2012 tests. In particular, we exploit the variation in the question ordering of this test to compute students' sustained performance throughout it. García-Pérez and Hidalgo-Hidalgo (2017) finds that PAE had a substantial positive effect on children's academic achievement and that a longer exposure to the program improves students' scores. Our study complements previous literature by focusing on the impact of this program on skills much more flexible among teenagers than test scores (Carneiro and Heckman, 2003; Almlund et al., 2011) and enriches the previous analysis with information on the number of students per school actually treated, allowing therefore to better identify the true effects. In addition, it adds

[2]For instance, Jacob (2002) and Bertrand and Pan (2013) show that girls have less disruptive behavioral problems and Cornwell et al. (2013) found that girls show more developed attitudes towards learning. See also Balart and Oosterveen (2019) and references therein.

[3]PAE is the Spanish acronym for Programa de Acompañamiento Escolar.

to works on non-self-assessed measures of non-cognitive skills by computing each student specific measure and analyzing whether a remedial intervention can improve these skills. Finally, it moves forwards the literature on gender gaps in education by studying whether non-cognitive returns to remedial education differ by gender.

We compare non-cognitive skills of students who attended schools that participated in the PAE with the hypothetical outcomes that these same students would have obtained had they not attended PAE schools. While doing so we have to cope with two difficulties. First, we cannot observe whether a particular student actually received the treatment, but only if she is in a treated school (intent-to-treat estimates). We provide a more precise estimate of the effect of the PAE by considering the number of students actually treated at school and the school as the treatment unit. The second challenge arises as schools participation into the PAE is not a random event which introduces two possible bias of opposite sign in the analysis. First, the fact that schools need to meet some eligibility criteria in order to participate into the PAE might introduce a negative bias. And second, a problem of self-selection emerges as schools volunteered for the program, which might introduce a positive bias. To tackle the first source of bias we implement a matching procedure weighting method. The counterfactual outcomes for students in treated schools are inferred using schools that did not participate in the PAE but took the PISA exams. And, to ensure that treatment and control groups are comparable on observables, students in the control group are re-weighted by assigning relatively more weight to those students whose individual, family and school characteristics are similar to students in the treated group.[4] To carefully address the positive bias we proceed as follows. First, the richness of our data allows us include a set of variables that capture these parents', teachers' and school's characteristics in the propensity score estimation that might help reduce the bias. And second, our access to schools' performance before some schools joined the program and after joining it allows us to estimate the impact of the program following a *difference-in-difference* approach.

We find that educational interventions aimed at teenagers can improve their non-

[4]See Lavy et al. (2020), García-Pérez and Hidalgo-Hidalgo (2017) or Hospido et al. (2015) who use a similar empirical strategy.

cognitive skills. In particular we show that the PAE has a substantial positive effect on our main measure of this type of skills: the estimated increase on the ability to sustain test performance is between 0.041 and 0.047 of one standard deviation. In addition, it reduces the probability of falling behind into the bottom part of the ability to sustain test performance distribution by about 2 percentage points. The corresponding figures for girls are 0.094 of one standard deviation and 4.4 percentage points. As treated schools in our sample participate in the PAE, on average, for three years, should the impact be the same for every year, then the impact of being treated one year on girls could be of 0.03 of one standard deviation and 1.47 percentage points, respectively. We found no statistically significant impact of the program on boys.

Such result is not due to a larger proportion of girls in the percentiles of the outcome distribution where the impact of the program is larger, nor to a higher participation of girls to it, or to gender differences in test taking strategies. It is plausibly explained by the fact that girls participate more intensively and they better respond to the remedial education activities. Our results hold when we consider the school as the unit of analysis. In addition, following a difference-in-difference approach we find similar results.

The paper is organized as follows. Section 2 provides a summary of the related literature and how this paper contributes to it. Section 3 presents our measure of sustained test performance. Section 4 summarizes the remedial program and presents data and descriptive statistics. Section 5 describes the methodology. Section 6 reports the baseline results of the impact of the intervention. Section 7 provides results of its possible heterogeneous effects and discusses the validity of these findings. Section 8 concludes.

2 Literature review

Our paper contributes to three strands of the literature: the evaluation of remedial education programs, the research on non-cognitive skills and the literature on gender differences in both cognitive and non-cognitive skills. The first strand of literature studies the impact of remedial education programs mostly on students' cognitive skills. Lavy and

Schlosser (2005) and Lavy et al. (2020) evaluate the short-term and long-term effects of the Bagrut 2001 program, a remedial intervention very close in spirit to the one proposed to be evaluated in this study, which provided additional instruction to underperforming high school students in Israel. Their results suggest that remedial education was more cost effective than alternatives based on financial incentives for pupils and teachers and that there are positive returns at adulthood, in terms of completed years of education and increased income mobility.[5] Importantly for our study, Heckman (2000) provides a review on several interventions in the nineties in the US that operate during the adolescent years. These programs were either mentoring type or incentive-based activities promoting non-cognitive skills oriented towards disadvantaged teenagers and were found to be effective. Therefore, he concludes that social policy should be more active in attempting to alter non-cognitive traits, especially in students from disadvantaged environments who receive poor discipline and little encouragement at home. Non-cognitive skills were also the objective of the remedial education programs studied by Holmlund and Silva (2014), Battaglia and Lebedinski (2015) and Martins (2017). The current paper departs from the previous works by studying the impact of this type of interventions on non-cognitive skills as measured by the ability to sustain the performance during the test. To the best of our knowledge, we provide novel evidence on the impact of educational interventions aimed at teenagers on non-cognitive skills. It also adds to García-Pérez and Hidalgo-Hidalgo (2017) in at least two ways. First, it better identifies the impact of the program using the number of treated students at school. Second, as PISA test scores might capture both cognitive and non-cognitive abilities, also depending on the length of the test, it thus focuses on a cleaner measure of students' skills: their non-cognitive abilities.

This paper therefore relates to recent works on non-self assessed measures of non-cognitive skills. Borghans and Schils (2018) use the ability to sustain test performance, that is the rate of decline in performance over the course of the 2006 PISA test's administration to measure non-cognitive factors such as agreeableness, motivation and ambition.

[5] A number of recent papers have focused on remedial programs in tertiary education in Europe and the US. For example, De Paola and Scoppa (2014, 2015) analyse the impact of remedial courses on the achievement of college students in Italy. Bettinger and Long (2009) and Calcagno and Long (2008) study the causal effect of remediation on the outcomes of college students in Ohio and Florida, respectively.

Using 2009 PISA, Zamarro et al. (2019) expand the methods used by Borghans and Schils (2018) and find that the decline in test performance is a good predictor of international variation in test scores. Balart et al. (2018) decomposes the performance on the PISA test into two components: the starting level and the decline in performance during the test. The authors find that countries differ in the starting level and in the decline in performance, and that these differences are stable over time and positive and statistically significant associated with economic growth. Our paper complements their research by computing each student specific rate of decline during test performance instead of focusing on an aggregate measure at country level. In addition it studies whether remedial education programs can help to improve these skills.

Finally, we contribute to the literature on gender gap in education. Gender gaps in cognitive skills have long been studied by economists. The main finding is that, on average, girls perform better than boys in reading tasks whereas boys outperform girls in maths and science tasks (see Fryer and Levitt (2010), Cornwell et al. (2013) or, more recently, Nollenberger et al. (2016) and references therein). Most closely related to our paper, Balart and Oosterveen (2019) considers gender differences in non-cognitive skills as measured by performance during the test, and finds that the relative performance of girls improves as the test proceeds. This result is in line with findings in the literature that suggest that girls tend to perform better than boys in several measures of non-cognitive skills. Our findings confirm these conclusions and move forward them by analyzing whether girls are not only better than boys in non-cognitive skills but also more apt to improve them when receiving remedial education. We therefore also relate to Bertrand and Pan (2013) which documents not only a gender gap in non-cognitive skills but also gender differences in the non-cognitive returns to external inputs.

3 Measuring non-cognitive skills

Non-cognitive skills usually refer to work and study habits, such as motivation and discipline, and behavioral attributes, such as self-esteem and confidence (ter Weel, 2008;

Holmlund and Silva, 2014). Often, such characteristics are self-assessed. Nevertheless, self-assessed measures might be biased by a lack of self-knowledge and subject to manipulation by students who can benefit from suggesting specific personality traits (see Sternberg et al. (2000), among others).

This evidence motivates the use of answering patterns to obtain measures of non-cognitive skills that do not rely on self-reports. We build on previous research (Balart and Oosterveen, 2019; Zamarro et al., 2019; Borghans and Schils, 2018; Rodríguez-Planas and Nollenberger, 2018) which uses students' response patterns to surveys and tests to get a non-self-assessed measure for their personality traits.[6] The idea is that students' test scores are not just the result of cognitive skills but also, and as doing the test takes time, of the ability to sustain performance throughout it. Students, through their effort on tests and surveys, provide information about their conscientiousness, self-control or persistence. Building up on this notion, Borghans and Schils (2018) and Balart et al. (2018) propose an approach to decompose students test scores into two elements: their initial performance and the decline in performance. The aim of this decomposition is precisely to capture both types of skills: whereas the initial performance provides a measure of cognitive skills, the performance decline is a measure of non-cognitive skills.[7] The analysis of the latter is the focus of this paper.[8] Of course, as Borghans et al. (2008) or Brunello et al. (2018) among others recognise, it is both conceptually and empirically very difficult to separate cognitive ability from non-cognitive skills. For instance, initial performance in a test might be influenced by non-cognitive abilities as motivation. To

[6]In Section A of the Online Supplementary Material we also comment on results for students' self-assessed measures such as absenteeism and truancy, discipline measured by the way students behave in class, self-confidence, sense of belonging to the school, and perception of learning at school.

[7]Borghans and Schils (2018) and Balart and Oosterveen (2019) provide several arguments in favor of this idea. For instance, the former find that students with higher levels of agreeableness (a Big Five personality trait), ambition and motivation towards learning have a smaller performance decline. In addition, they show that the performance decline predicts future outcomes above and beyond the pure test score. Balart and Oosterveen (2019) argued that if the performance decline were in fact induced by cognitive skills, then we should observe girls experiencing a less pronounced decline in reading, while boys would have a less pronounced decline when answering math and science questions. They indeed typically score better on these subjects and girls on reading. However, they found the opposite: girls exhibited a less pronounced decline than boys in both reading and in math/science.

[8]The comparison of the impact of the PAE on both cognitive and non-cognitive skills is out of the scope of this paper. Nevertheless, to complement our results on non-cognitive skills here, we also analyse the impact of the program on students' initial performance and final score. See Section B of the Online Supplementary Material.

the extent that this is true, then by estimating the impact of the program on students' ability to sustain test performance we are underestimating its impact on non-cognitive skills.

Following recent literature, we therefore exploit the variation in the question ordering of a test to define our measure of non-cognitive skills: a student's sustained test performance. We computed it as the decline in performance throughout the PISA test, controlling for initial performance.[9] We use microdata on each students' answer to every single administered question in PISA 2012 for Spain. Using both the codebooks and information provided by the OECD, we retrieve which question the student had to answer on each position of the test. As also acknowledged in the related literature, PISA tests have two characteristics that are crucial for investigating student's differences in performance during the test. First, PISA uses multiple test booklets with different orders for different subjects. Each booklet can contain four different clusters in three different subjects: maths, reading and science. Second, these booklets are randomly assigned to students (OECD, 2013). This random assignment ensures that the variation in question numbers, that results from the ordering of clusters, is unrelated to characteristics of students.

Table 1: Rotation design of the 13 PISA booklets

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4	# q	# q Math	# q Reading	# q Science	# Students	# Girls	# Boys
1	Math 5	Science 3	Math 6A	Science 2	60	25	-	35	875	457	418
2	Science 3	Reading 3	Math 7A	Reading 2	58	12	29	17	872	446	426
3	Reading 3	Math 6A	Science 1	Math 3	57	25	14	18	884	426	458
4	Math 6A	Math 7A	Reading 1	Math 4	52	37	15	-	871	445	426
5	Math 7A	Science 1	Math 1	Math 5	54	36	-	18	859	436	423
6	Math 1	Math 2	Reading 2	Math 6A	51	36	15	-	864	437	427
7	Math 2	Science 2	Math 3	Math 7A	53	35	-	18	875	454	421
8	Science 2	Reading 2	Math 4	Science 1	63	12	15	36	864	432	432
9	Reading 2	Math 3	Math 5	Reading 1	54	24	30	-	881	427	454
10	Math 3	Math 4	Science 3	Math 1	53	36	-	17	826	404	422
11	Math 4	Math 5	Reading 3	Math 2	49	35	14	-	822	424	398
12	Science 1	Reading 1	Math 2	Science 3	61	11	15	35	813	415	398
13	Reading 1	Math 1	Science 2	Reading 3	59	12	29	18	819	418	401

Source: PISA 2012

As shown in Table 1, PISA 2012 has 13 different versions of the test (booklets), all of them containing four clusters of questions q (test items). A booklet contains approximately 50 to 60 test items. Each cluster of questions takes 30 minutes of test time and

[9]The PISA test is intended to evaluate educational systems and to provide comparable data by measuring 15-year-old school pupils' scholastic performance on mathematics, science, and reading. It is a worldwide study by the Organisation for Economic Co-operation and Development (OECD). It was first performed in 2000 and then repeated every three years.

students are allowed a short break after one hour. Clusters labeled *Math 1* to *Math 7A* denote the seven paper-based standard mathematics clusters, *Reading 1* to *Reading 3* denote the paper-based reading clusters, and *Science 1* to *Science 3* denote the paper-based science clusters.[10] Each cluster appears in each of the four possible positions within a booklet once (OECD, 2013). This means that one specific test item appears in four different positions of four different booklets and, since clusters have different number of items, each question has a different position in each test. For instance, cluster *Maths 5* is included in booklets 1, 5, 9 and 11 as respectively the first, forth, third and second cluster. As it can be observed, the number of students that took each booklet is very similar and ranges from 813 to 884. Note also that each booklet is almost evenly shared by boys and girls. To construct our measure of student's *individual* rate of decline in test performance, we estimate the following specification for *each* student i:

$$Pr(y_{qi}) = \Phi(\alpha_0 + \alpha_1 p_{qi} + \alpha_2 d_q + u_{qi}), \tag{1}$$

where y_{qi} is a dummy for whether student i answered question q correctly, Φ is the standard cumulative normal distribution, p_{qi} is the position of question q in the version of the test answered by student i and it is rescaled such that the first question is numbered as 0 and the last question as 1 and d_q is a binary variable capturing the difficulty of the question q (it is equal to 1 for multiple choice or open, and 0 for simple choice questions).[11]

Our coefficient of interest is α_1 which shows the individual pattern of the test performance (recall that we have one α_1 for each student). A significant and negative (positive)

[10] Balart and Oosterveen (2019) compare students' performance in the standard paper and pencil tests used in most PISA exams and the PISA 2015 test which was given on the computer and navigation across question units was restricted. The authors find no differences in students' test behaviors.

[11] We tested, student by student, the correlation between the position of the question and its difficulty, finding a negligible correlation coefficient. We also provide an additional measure of difficulty in line with Item Analysis Statistics: the percentage of students who correctly answer the question. This Item Difficulty Index ranges from 0 to 100; the higher the value, the easier the question (Lord, 1952). Note that, within a single booklet, question position could turn out to be correlated with question characteristics (for instance, whether maths questions appears only at the beginning of the booklet). However, we find that this is not the case here. As an additional check, we also computed the pattern of performance by estimating equation (1) for each school, which allows us to add booklet-specific fixed effects. We found that this measure is not statistically different from the one used throughout the main text. Finally, as an alternative definition of correct answer, we recode a question as correct if the answer is correct or partially correct. See Section C of the Online Supplementary Material for comments on robustness of our main results to these alternative definitions and checks.

coefficient would reveal a decline (improvement) in performance from the first to the last question of the test.[12] As our dependent variable is a dummy, we estimate a probit model.[13]

Figure 1 depicts the distribution of the pattern in performance (that is, the marginal effect of p_q) during the test. As it can be observed, the majority of the students shows a decline in performance, in line with previous findings by Borghans and Schils (2018), Balart and Oosterveen (2019) or Zamarro et al. (2019). In our setting, about 65% of students performs worst towards the end of the test than at the beginning. The average estimated pattern in performance is negative, precisely it is equal to -0.097, which means that the probability to answer the last question correctly is 9.7 percentage points lower than the probability to answer the first question correctly.[14] Therefore, from now on we refer to α_1 as the individual rate of decline.[15]

Figure 2 reports average estimated rate of decline as in Equation (1) separately for boys and girls. The average estimated rate of decline is lower among girls, which is also in line with recent evidence by Balart and Oosterveen (2019). As it can be observed, there is an initial gap in test scores favoring boys, however, during the test this advantage vanishes and girls finish the questionnaire outperforming boys.[16]

In the rest of the paper we focus on the student's decline in test performance and consider the following two outcome variables: (i) each student's rate of decline; (ii) the probability of falling behind the general progress of the group in terms of rate decline, that is being in the first quartile of the rate of decline distribution.

[12]Balart and Oosterveen (2019) also check for the non-linearity effect of the position of the question finding similar qualitative results than under the linear assumption.

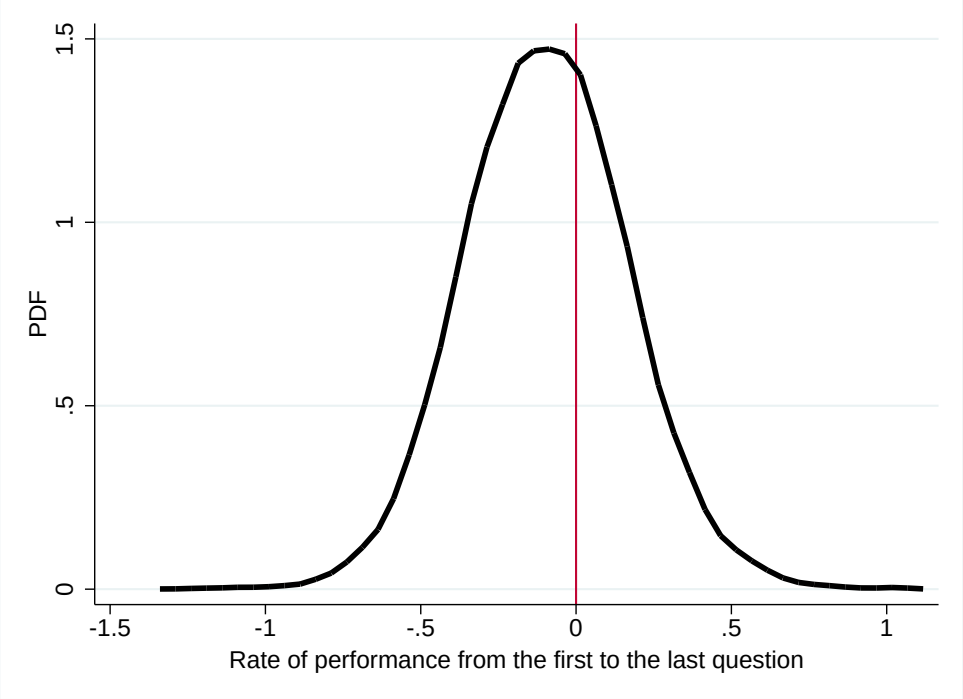
[13]In line with Hitt et al. (2016) and Zamarro et al. (2018) we also consider the number of items reached during the test as an alternative measure of student's non-cognitive skills. Results on this measure are commented in Section D of the Online Supplementary Material.

[14]The drop in the percentage of correct answers from the first to the last cluster (56.23%, 55.55%, 54.06% and 51.84% for the first, second, third and fourth cluster respectively) constitutes additional evidence supporting the decline in performance during the test.

[15]As expected, on average the initial performance is larger among students with a rate of decline throughout the test than among those with a rate of incline. Therefore, we control for students' initial performance when estimating the impact of the program on their rate of decline.

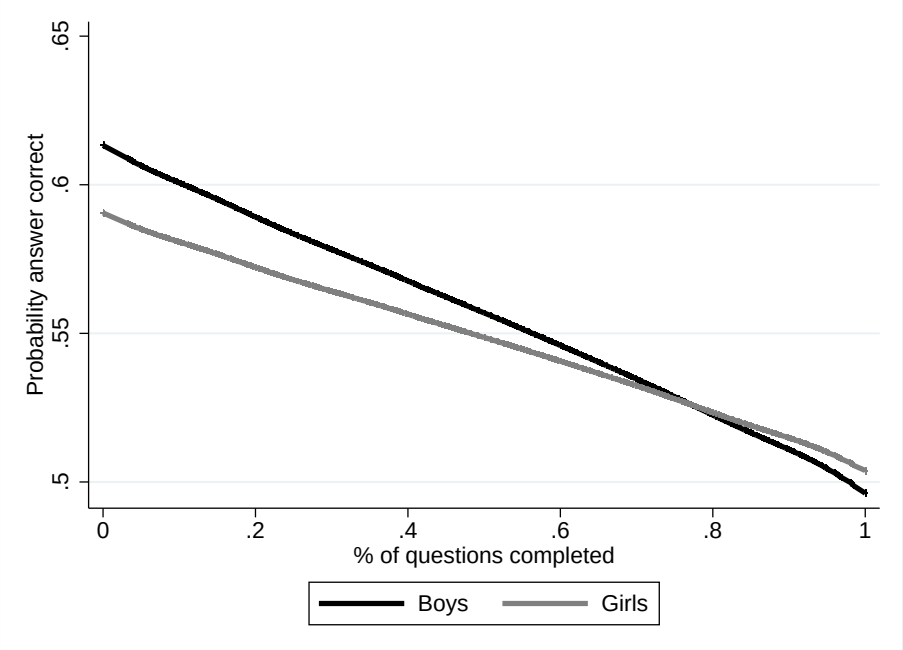
[16]Gender difference is statistically significant at 0.01 level. A similar pattern of performance is observed by considering the order of subjects, that is, whether maths is taken before reading and vice versa, although the gender gap is lower among students with assigned booklets where maths questions appear before the reading ones (see Section C in the Online Supplementary Material).

Figure 1: Distribution of estimated pattern in performance



Note: PDF of the students pattern of performance. Negative (resp. positive) values shows decline (improvement) in performance from the first to the last question of the PISA test.

Figure 2: Estimated pattern in performance by gender



Note: The figure uses the Stata command LOWESS to visualize the relationship between the probability to answer a question correctly and the position of the question as in Equation (1) for both boys (black) and girls (grey) with a default bandwidth of 0.8N.

4 The remedial program

The Program for School Guidance (PAE) is a program targeting public primary and secondary schools. The aim of this intervention was to enhance the learning abilities and academic returns of underperforming students with poor socioeconomic backgrounds. It consisted of providing support (at least 4 hours per week) during after-school hours to those students with special needs and learning difficulties. During the remedial classes, the students engaged in guided reading and worked on the subjects that presented particular difficulties for them. Instructors offered clarification, provided additional material and assisted students with work organization techniques. An important objective was to improve their social abilities and studying techniques. Therefore this type of intervention is expected to have positive impacts on students' skills such as motivation, persistence and self-control which are captured in our non-cognitive ability measure. It is more about study habits and behavioral attributes, than the content of the subject. The support was provided in small groups of on average 5-10 students by instructors or teachers from the students' own schools. Students were selected by both their tutor and the rest of the teachers and could be in any grade within the school. They were chosen based on their poor academic results, general motivation and prospects, although there was no single quantifiable and explicit selection rule.

The intervention was jointly financed by both the central and the regional governments. The criteria to distribute funds for the program among regions included the number of public schools, the number of students attending public schools and the number of early school leavers or dropouts. Schools volunteered for the program. Unfortunately there is not an explicit percentage threshold of students from poor background required for the school to be admitted to the program. Nevertheless, apparently, the guidelines to distribute funds among schools within regions resemble the previous iterations: big schools with a high number of early leavers and dropouts were more likely to participate in the program.

The PAE was progressively introduced throughout the period 2005-2012. Even though it was implemented in both primary and secondary schools, we focus our analysis on

secondary schools because PISA 2012 exam is taken by 15-year-old students. We consider the last four academic years the program was in place, that is, from 2008 till 2012, when students in our sample were in grades 7 to 10 and were attending the same secondary schools where they took the PISA exams (10th grade). Note that, even if the secondary schools participated in the program during the 2005-2008 period, students in our sample did not benefit from it since they were still attending primary school.[17]

As the program was implemented only in public schools, we exclude from the sample both private and private but publicly financed schools. In addition, we do not consider in the analysis schools that joined other remedial programs or where the PAE was implemented during any academic year between 2005/06 and 2010/11 but not in 2011/12, the year when the PISA exams were taken.[18] Our sample consists of 11,105 individuals from 395 schools.[19]

We consider as *treated* those students at schools that participated in the PAE during the same academic year in which PISA exams were taken, namely, 2011/12, regardless of whether the school joined the program before (that is, in any academic year between 2008/09 and 2010/11). Our treated schools have participated in the program, on average, for three out of four years.[20] We consider as *controls* students in schools where the PAE was not implemented at all (that is, in no academic year between 2008/09 and 2011/12). As a result, there are 130 treated schools (with 3,694 students) and 265 control schools (with 7,411 students) in our sample.

[17]The Spanish education system is organized into three levels: primary (grades 1-6), secondary (grades 7-10) and pre-college (grades 11-12). The first two levels are compulsory (a student can choose to leave school at age 16). School starts at 6 years old. Most schools provide either primary or secondary and pre-college education. See Spanish Ministry of Education (2016).

[18]We excluded 352 schools because they are private or private publicly financed schools. From the remaining 550 public schools in PISA 2012 database, we exclude 133 schools because they participated in other remedial programs and 22 schools where the PAE was implemented only before the academic year 2011/12.

[19]See the PISA 2012 Technical Report for details on PISA 2012 and García-Pérez and Hidalgo-Hidalgo (2017) for details on how the PAE was introduced in the schools.

[20]Alternatively we could analyse the effect of the program considering as treated those students in schools implementing the program for the first time in the academic year 2011/12. The low number of treated schools according to this definition (only 17) impedes us from using the specification for the propensity score estimation adopted in the rest of estimations in the paper and thus results are not completely comparable.

4.1 Students' Characteristics

The PISA 2012 database provides microdata on each student's answer to each question, individual-level information on demographics (e.g., gender, immigration status, month and year of birth), socioeconomic background (parental education and occupation) and school-level variables. Table 2 reports the main descriptive statistics of a set of individual, socioeconomic and school-level variables in our sample (in column (1)). It also reports descriptive statistics for students in treated schools (column (2)), students in control schools (column (3)) and the differences between them (column (4)).

There are no statistically significant differences with respect to gender composition between the two groups. However, students in PAE schools differ from those in schools that did not join the program: control students are less likely to be migrants and are less likely to have repeated a grade. In addition, the proportion of educated parents and the index of educational materials are lower among students in treated schools, suggesting that these schools have a higher proportion of students from disadvantaged backgrounds. Students in treated and control schools also differ in their initial test score, measured as the average test score in the first five questions of the first cluster of the test, which is statistically significantly lower among the former.[21] Finally, treated schools are larger and exhibit a larger proportion of dropouts and lower Economic Social Cultural Status (ESCS). Conversely, control schools have a higher student-teacher ratio, parents exert less pressure on teachers and teachers contribute to a higher extent to create good school climate.

Table 3 shows the (standardized) estimated rate of performance for the complete PISA test and the first quartile (P25), that is a dummy variable equal to 1 for students with rate of performance in the first quartile of the rate of performance distribution of the sample. Table 3 reports the values overall and by gender for the all sample and for the treated and control group. A negative (positive) value measures the % reduction (increase) in

[21] We face a trade-off when selecting the number of questions considered as initial score. The larger is the number of questions, the lower the number of missing values for the initial score as students may jump some initial questions in the test. However, including many questions makes more difficult to assume that *initial score* is not capturing non-cognitive skills. As an approximation, we consider the first five questions.

Table 2: Summary statistics

Variable	(1) All	(2) Treated	(3) Controls	(4) P-value Diff. (2)-(3)
Individual variables				
Initial test score[a]	.621	.607	.628	.000
	(.272)	(.274)	(.271)	
Girl(=1)	.506	.499	.509	.312
	(.5)	(.5)	(.5)	
Migrant(=1)	.107	.149	.086	.000
	(.309)	(.357)	(.281)	
Repeated once(=1)	.237	.272	.22	.000
	(.425)	(.445)	(.414)	
Repeated more than once(=1)	.087	.106	.078	.000
	(.282)	(.308)	(.268)	
Attended kindergarden(=1)	.839	.83	.844	.054
	(.367)	(.376)	(.363)	
Socioeconomic variables				
Index of education possession[b]	.063	.041	.074	.068
	(.885)	(.887)	(.885)	
Mother highly educated(=1)[c]	.345	.303	.365	.000
	(.475)	(.46)	(.482)	
Father highly educated(=1)[d]	.33	.299	.346	.000
	(.47)	(.458)	(.476)	
School variables				
School size (no. students)	606.893	621.813	599.512	.000
	(318.336)	(278.108)	(336.328)	
Prop. of dropout	.095	.115	.085	.000
	(.109)	(.111)	(.107)	
Prob. of dropout in high quartile(=1)	.236	.308	.2	.000
	(.425)	(.462)	(.4)	
ESCS[e]	-.274	-.371	-.224	.000
	(.977)	(.971)	(.974)	
ESCS in high quartile(=1)	.272	.156	.329	.000
	(.445)	(.363)	(.47)	
Student-Teacher Ratio	9.621	9.264	9.798	.000
	(7.213)	(2.052)	(8.704)	
Parental pressure on teachers(=1)[f]	.356	.391	.338	.000
	(.479)	(.478)	(.473)	
School climate-teacher(=1)[g]	.564	.686	.504	.000
	(.496)	(.464)	(.5)	
Rural(=1)[h]	.42	.411	.424	.194
	(.494)	(.492)	(.494)	
Observations	11,105	3,694	7,411	

Standard deviations in parentheses.

[a] Initial test score corresponds to the average score in the first five questions of the first cluster of the test.

[b] The index of education possession indicates whether the home possesses a desk and a quiet place to study, a computer and/or educational software and books to help with school work, and a dictionary. It ranges between -3.93 and 1.12.

[c] The mother is defined as highly educated if she has achieved at least tertiary education.

[d] The father is defined as highly educated if he has achieved at least tertiary education.

[e] Index of economic, social, and cultural status.

[f] The dummy is equal to 1 if the principal claims that parents exert pressure into teachers and principal to improve the school quality.

[g] It is a dummy equal to 1 if the school is below the median value of the index of teacher-related factors affecting school climate. Positive values indicate that the teacher-related behaviors hinder learning to a lesser extent. The index ranges between -3.2778 + 2.8533.

[h] It is a dummy equal to 1 if the school is located in a village or a small town.

the probability of correctly answering a question as the position of that question increase 1% from the first to the last question.

Table 3: Students' outcomes:
Rate of performance from the first to the last question

	Boys			Girls			All students		
	All	Treated	Control	All	Treated	Control	All	Treated	Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mean	-.061	-.091	-.046	.06	.12	.03	0	.014	-.007
	(1.013)	(1.033)	(1.002)	(.984)	(.964)	(.992)	(1)	(1.005)	(.998)
First quartile (P25)	.276	.29	.269	.224	.197	.238	.25	.243	.253
Observations	5,430	1,843	3,587	5,581	1,841	3,740	11,011	3,684	7,327

Standard deviations in parentheses. The rate of performance is standardized with the average and standard deviation of the sample of students.

On average, we observe an increase in the probability of correctly answering a question as the position of that question increase for treated schools and for girls, while boys are more likely to answer badly. The percentage of boys in the first quartile of the rate decline distribution is larger than the percentage of girls (0.276 vs 0.224).

5 The empirical strategy

We study the effects of the PAE on the students' rate of decline in test performance and on their probability of falling behind the general progress of the group (having a rate decline in the first quartile of the rate decline distribution). Our estimates are intention-to-treat estimates: we cannot observe whether a particular student actually received the treatment, but only if she is in a treated school. To provide a more precise estimate of the effect of the PAE we exploit the information on the number of students actually treated in each school and also consider the school as the treatment unit.

Recall that schools' participation into the PAE is not a random event. The fact that schools need to meet some eligibility criteria in order to participate into the PAE might introduce a negative bias to the extent that schools are required an important number of students from disadvantaged backgrounds and poor academic results (i.e. a vulnerable student body). Observe that, eligibility criteria, even though ambiguous, is based on observable characteristics by the policy maker and also by us through the extensive questionnaire provided by both parents and schools in our sample. By considering these

characteristics (parents' socioeconomic status, percentage of school dropouts, etc.), we can re-weight the sample of students in control schools such that they can provide a counterfactual to the sample of students in treated schools. If D_i denotes a binary variable that indicates exposure to the treatment, then this is defined as the probability of PAE participation conditional on some pre-treatment characteristics, X_i:

$$p(X_i) \equiv Pr(D_i = 1|X_i) = E(D|X_i). \tag{2}$$

Now, let Y_i^1 denote the potential outcome that student i would have obtained had she received the PAE treatment and Y_i^0 had she not received the PAE treatment. Therefore, the average effect we are interested in estimating when evaluating the PAE is

$$\tau = E(Y_i^1|D = 1) - E(Y_i^0|D = 1). \tag{3}$$

The second term in the equation above is the counterfactual outcome in the absence of the treatment and thus is unobservable and must be estimated. This is achieved by using the outcomes of control students, that is, students in schools where the PAE was not implemented at all. It requires the characteristics of the control and treatment group to be as similar as possible. In our sample, as previously mentioned, students in treated and control schools differ in their demographic characteristics and in socioeconomic background (see Table 2). To address this problem, we use information on demographic, parental and school characteristics in the PISA 2012 database to *re-weight* the sample of controls such that they can provide a counterfactual to the PISA outcomes of students in treated schools. Under the standard assumptions of conditional independence or unconfoundedness and common support, we have that:[22]

$$E(Y_i^0|D = 1) \equiv E(\omega(x_i)Y_i|D = 0), \tag{4}$$

[22]The assumption of conditional independence or unconfoundedness requires that within each cell defined by X_i, treatment is random, or similarly, the selection into treatment depends only on the observables X_i. The common support assumption, $p(X_i) \in (0, 1)$, can be tested by comparing the propensity score densities of the treated and control groups. We check it graphically in Figure E.1 in the Online Supplementary Material. As it can be observed, the common support assumption holds in our sample.

where $\omega(x_i) = \frac{1-\pi}{\pi} \times \frac{p(X_i)}{1-p(X_i)}$ and $\pi = \Pr(D_i = 1)$.

Observe that the weights, $\omega(x_i)$, increase the relevance in the control sample of those individuals who are very similar to students in treated schools, where similarity is defined here by the predicted probability of participation in a logit that explains participation given pre-treatment characteristics, that is, by the propensity score, $p(X_i)$.[23] We therefore compute the *inverse probability weighting estimator (IPWE)* by regressing the outcome variables on the treatment, where each observation is weighted by $\omega(x_i)$.[24] Since, through the consideration of the propensity score in the weighting procedure, there is a control for all covariates, X_i, in this estimation there is no need to include them. In any case, we may also include the covariates, X_i, as a robustness check.

Column (3) of Table 4 presents the means of the control sample once the latter is re-weighted by $\omega(x_i)$.[25] Column (4) reports the differences in characteristics between treated and re-weighted controls. As expected, these are not statistically different from one another, particularly for the set of controls considered in the propensity score estimation (i.e., the balancing property is satisfied). Finally, note that the sample is also similar along characteristics that we do not include in the propensity score (ESCS and father's education). The similar composition of treated and re-weighted control groups even in characteristics omitted from the propensity score reinforces the credibility of the assumption that treated and re-weighted control students would have performed similarly had the treated students not been treated (see Lavy and Schlosser (2005) or Hospido et al. (2015) for a similar check).

Furthermore, a problem of self-selection emerges as schools volunteered for the program, which might introduce a positive bias. Program participation could be due to, among other reasons, special interest by parents, teachers or school principals. This implies

[23]The estimates of the probability of participation in the remedial program are provided in Table F.1. of the Online Supplementary Material.

[24]See Lavy et al. (2020), García-Pérez and Hidalgo-Hidalgo (2017) and Hospido et al. (2015) for a similar empirical strategy. See also Abadie and Imbens (2002) for details regarding the use of OLS with the matching procedure weighting and Hirano et al. (2003) or Busso et al. (2014) for further methodological details.

[25]For those observations with missing values for some of the variables included in the propensity score the estimated propensity score will be missing and, thus the weight variable will be missing too. This explains the difference between the controls observations in column (3) in Table 2 (7,441) and the weighted controls observations in column (3) in this table (7,331).

Table 4: Summary statistics re-weighted

Variable	(1) All	(2) Treated	(3) Weighted Controls	(4) P-value Diff. (2)-(3)	(5) P-score
Individual variables					
Initial test score[a]	.621	.607	.606	.928	yes
	(.272)	(.274)	(.276)		
Girl(=1)	.506	.499	.496	.772	yes
	(.5)	(.5)	(.5)		
Migrant(=1)	.107	.149	.16	.165	yes
	(.309)	(.357)	(.366)		
Repeated once(=1)	.237	.272	.272	.982	yes
	(.425)	(.445)	(.445)		
Repeated more than once(=1)	.087	.106	.114	.199	yes
	(.282)	(.308)	(.318)		
Attended kindergarden(=1)	.839	.83	.827	.722	yes
	(.367)	(.376)	(.378)		
Socioeconomic variables					
Index of education possession[b]	.063	.041	.056	.428	yes
	(.885)	(.887)	(.894)		
Mother highly educated(=1)[c]	.345	.303	.303	.921	yes
	(.475)	(.46)	(.459)		
Father highly educated(=1)[d]	.33	.299	.3	.886	no
	(.47)	(.458)	(.458)		
School variables					
School size (no. students)	606.893	621.813	625.856	.472	yes
	(318.336)	(278.108)	(278.784)		
Prop. of dropout	.095	.115	.118	.177	yes
	(.109)	(.111)	(.118)		
Prob. of dropout in	.236	.308	.304	.714	yes
high quartile(=1)	(.425)	(.462)	(.46)		
ESCS[e]	-.274	-.371	-.372	.957	no
	(.977)	(.971)	(.953)		
ESCS in high quartile(=1)	.272	.156	.158	.868	yes
	(.445)	(.363)	(.364)		
Student-Teacher Ratio	9.621	9.264	9.258	.909	yes
	(7.213)	(2.052)	(2.882)		
Parental pressure on teachers(=1)[f]	.356	.391	.398	.475	yes
	(.479)	(.478)	(.49)		
School climate-teacher(=1)[g]	.564	.686	.692	.525	yes
	(.496)	(.464)	(.462)		
Rural(=1)[h]	.42	.411	.422	.274	no
	(.494)	(.492)	(.494)		
Observations	11,105	3,694	7,331		

Standard deviations in parentheses.

[a] Initial test score corresponds to the average score in the first five questions of the first cluster of the test.

[b] The index of education possession indicates whether the home possesses a desk and a quiet place to study, a computer and/or educational software and books to help with school work, and a dictionary. It ranges between -3.93 and 1.12.

[c] The mother is defined as highly educated if she has achieved at least tertiary education.

[d] The father is defined as highly educated if he has achieved at least tertiary education.

[e] Index of economic, social, and cultural status.

[f] The dummy is equal to 1 if the principal claims that parents exert pressure into teachers and principal to improve the school quality.

[g] It is a dummy equal to 1 if the school is below the median value of the index of teacher-related factors affecting school climate. Positive values indicate that the teacher-related behaviors hinder learning to a lesser extent. The index ranges between -3.2778 + 2.8533.

[h] It is a dummy equal to 1 if the school is located in a village or a small town.

that the conditional independence assumption is not satisfied. In order to deal with this issue we proceed as follows. First, we include a set of variables that capture these parents', teachers' and school's characteristics in the propensity score estimation that might help reduce the bias. In particular we consider the school ESCS, whether teachers foster good school climate and whether principal claims that parents exert pressure into teachers and principal to improve the school quality which capture the degree of both parents and teachers' commitment and motivation.[26] Second, our access to schools' performance in 2009 (before some schools joined the program) and in 2012 (after joining it) allows us to estimate the impact of the program following a *difference-in-difference* approach (see Section 6 below). By changing the conditional independence for the parallel trend assumption we can partially control for those unobserved school's and teachers' characteristics. Lastly, to assess the extent to which unobservables may drive our results, we follow Altonji et al. (2005) and Oster (2019) in calculating how strong selection on unobservables would have to be in order to explain the full observed relationship between the implementation of the program and the rate of decline. We find that the impact of unobserved factors would have to be at least fourteen times stronger, as compared to observed factors, in order to explain away the relationship between the program and the rate of decline. This makes it unlikely that unobservable factors can account for our results.[27]

Finally, in order to test our second hypothesis regarding the existence of heterogeneous effects of the program depending on students' gender we add an interaction term for the treatment and student's gender.

[26]In particular we consider the school index of economic, social, and cultural status, whether the principal claims that parents exert pressure into teachers and principal to improve school quality and whether the school is below the median value of the index of teacher-related factors affecting school climate.

[27]More precisely, if we look at our results with OLS and IPWE we observe that (i) with OLS, the implied ratios are negative: the observable controls are on average negatively correlated with the rate of decline, yielding stronger coefficient estimates than in the basic regression without controls. In these cases, the Altonji et al. (2005) test suggests that our OLS estimates are likely to be downward biased, provided that the unobservables are positively correlated with the observables. (ii) With IPWE we observe a positive correlation between the implementation of the program and the rate of decline whose ratios are 13.9 and 14. This implies that selection on unobservables would have to be at least fourteen times stronger than selection on observables for our main result to be overturned.

6 The baseline impact of the program

Panel A of Table 5 presents the estimated overall effect of the program on students' rate of decline and on their probability of belonging to the first quartile in the rate of decline distribution. The first two columns, and as a benchmark, show the results of a simple OLS estimation without and with covariates. The estimated coefficient in the two cases is not significant. However, recall that this approach produces estimates without taking into account that treated and control students differ in characteristics other than the treatment which, in turn, also affect their probability of being treated. The third column shows the re-weighting estimate without covariates. The observed impact of the program is 0.047 of one standard deviation and is statistically significant at the 5% confidence level. The effect is very similar (0.041) when we include all of the variables considered in the logit model used to obtain the weights and it is statistically significant at the 10% confidence level. The robustness of this result suggests that the specification of the model that predicts PAE participation is appropriate.

Recall that treated schools in our sample participate in the PAE, on average, for three years. Since the estimates above shows that the impact of participating during the period is equal to 0.041 of one standard deviation, should the impact be the same for every year, then the impact of being treated one year could be of 0.014 of one standard deviation.

Results in rows (3) and (4) show the estimated impact of the treatment on the probability of belonging to the first quartile in the rate of decline distribution. Again the first two columns present the result from a simple OLS model without and with covariates; columns (3) and (4) presents results using re-weighting estimates. The results are the same when re-weighting estimates are used without and with covariates and are consistent with previous findings. The program reduced the probability of belonging to the bottom quartile in the rate of decline distribution by about 2 percentage points.

The results presented above are ITT estimates: we are assuming that all of the students in schools with the PAE are treated, while some of them might not have received remedial education at all. By doing so, we are underestimating the impact of the PAE. Nonetheless, we might be capturing peer effects of treated on non-treated students and overestimate

Table 5: The impact of PAE on the rate of decline

	OLS		IPWE	
	(1)	(2)	(3)	(4)
Panel A: All sample				
			Level	
PAE	0.021	0.034	0.047**	0.041*
	(0.022)	(0.022)	(0.024)	(0.023)
Mean in control[a]	-0.007	-0.007	-0.035	-0.035
			P25 of the entire sample	
PAE	-0.010	-0.015	-0.020**	-0.019**
	(0.009)	(0.009)	(0.010)	(0.010)
Mean in control[a]	0.253	0.253	0.264	0.264
Controls	No	Yes	No	Yes
Observations	11,089	10,964	11,011	10,964
Panel B: Number of treated students in PAE schools				
Fewer Students than the median				
			Level	
PAE	0.019	0.029	0.034	0.031
	(0.028)	(0.029)	(0.029)	(0.029)
			P25 of the entire sample	
PAE	-0.014	-0.018	-0.018	-0.019
	(0.012)	(0.013)	(0.012)	(0.012)
Controls	No	Yes	No	Yes
Observations	8,986	8,888	8,908	8,888
More Students than the median				
			Level	
PAE	0.071**	0.093***	0.098**	0.096**
	(0.035)	(0.036)	(0.039)	(0.037)
			P25 of the entire sample	
PAE	-0.030**	-0.037***	-0.038**	-0.039***
	(0.014)	(0.014)	(0.015)	(0.015)
Controls	No	Yes	No	Yes
Observations	8,519	8,425	8,441	8,425
Test of equality of the coefficients				
Level	no	no	reject	reject
P25 of the entire sample	no	no	reject	reject

Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

[a] In columns (1) and (2) we report the mean of the control group. In columns (3) and (4) we report the mean of the weighted control group.

the impact of the PAE on treated students. In order to address these concerns, we, first, consider the actual number of treated students at school and, second, we replicate the analysis considering the school as the unit of observation.[28]

On average, in each treated school there are 30 students who receive remedial education support in each academic year from 2008 to 2012, with a relative high standard deviation: the number of students goes from 9 to 99. The average amount corresponds to 6% of the students in those schools, with some schools having 1% and others 45% of students treated.[29]

Panel B of Table 5 provides results for the impact of the program on our outcomes of interest depending on the number of treated students in PAE schools. First, we compare students in schools whose number of treated students is lower than its sample median to students in control schools (top panel). Second, we compare students in schools whose number of treated students is higher than its sample median to students in control schools (bottom panel).[30] In schools with a large number of treated students, the estimated impact of the program on the rate of decline is an increase of 0.098 of one standard deviation, which is more than double the impact on the full sample of schools (around 0.041). The probability of belonging to the bottom quartile is reduced by 3.8 p.p., so again double the effect of the full sample of schools. No impact of the program is observed for the sample of schools with a number of treated students lower than the median.

Two concerns might arise here. The first one is related to the possibility that differences in other characteristics between schools with many and few treated students might explain the result we observe. Nonetheless, their characteristics suggest that, if any, schools with

[28]We also analyse sub-samples of students in schools with higher probability of being treated according to the characteristics affecting the propensity score. First, we consider treated students at schools where the proportion of migrants or non-educated parents is above its sample median. Second, we consider treated students at schools where the proportion of migrants or non-educated parents is below its sample median. By doing so, we first increase and then reduce the likelihood that they actually participated in the program. Results, in line with the ones presented here, are reported in Section G in the Online Supplementary Material.

[29]Recent data on a very similar remedial program implemented after 2012 in the Region of Madrid reveal a larger rate of participation, around 10%. Note that the percentage could be even larger here: we considered as treated students at schools that participated in the program in 2011/12 regardless of whether they participated before (in the previous three years). Thus, should the group of treated students during one academic year not fully overlap the group of treated students in the rest of the period, then the aggregate participation rate in our sample could be larger than 6%.

[30]We estimated the propensity score separately for each subsample and re-weighted them.

many treated students should perform worse than the ones with a lower number of them.[31] Second, we do not have information for the whole sample of treated schools: only 65% of them provided some data on the implementation of the program. Consolingly, students in schools sharing the information are comparable to students in schools not sharing them for most of the characteristics we use in the analysis (Section H in the Online Supplementary Material).[32]

Lastly, we replicate the same analysis as above but we consider the school as the unit of analysis, instead of the student. Results are in line to those at the student level, see Section I in the Supplementary Material.

6.1 Alternative approach: difference-in-difference

Finally we propose to use an alternative econometric approach which consists on a difference-in-difference model. In order to do so we consider two time periods, t: 2009 (pre-treatment) and 2012 (post-treatment). We do not use the differences-in-differences method as our main empirical strategy due to an important reduction in sample size. The composition of the sample may change between periods (note that students observed in 2009 are not the same as those observed in 2012 and the same is true for all pre-treatment years) which may confound any difference-in-difference estimate whenever the effect is attributable to that change in the population. Recall that the PISA school sample fluctuates at each wave: some schools stay in but some others do not. Nevertheless, next we show that our results are robust to the use if this alternative approach.[33]

Results for the differences-in-differences analysis are reported in Table 6 below.

For both outcomes we estimate the regressions without and with controls (columns

[31] As a complementary analysis, we consider the number of students involved in the remedial program per teacher and/or monitor. Both teachers from the own school and monitors provided support to the students in the program: on average, there are roughly 3 teachers and 2 monitors performing the remedial activities and they count for the 4% of all teachers in that school. This implies that during remedial education classes there are on average 9 students per teacher, our PAE student-teacher ratio. Results show that students at those schools where the program was better implemented (lower ratio) reassuringly benefit more from it, especially if they belong to the lower quartile of the distribution (see Table H.2 in Section H in the Online Supplementary Material).

[32] In schools providing information on the implementation of the PAE there are slightly more female students who did not attend kindergarden and who belongs to the upper quartile of the ESCS distribution.

[33] More details on the specification and on the differences between the two samples are reported in Section J of the Online Supplementary Material.

Table 6: The impact of PAE on the rate of decline (Diff-in-Diff)

	Level		P25 of the entire sample	
	(1)	(2)	(3)	(4)
Panel A: Difference-in-Difference				
PAE	0.008	-0.033	-0.020	-0.004
	(0.037)	(0.037)	(0.018)	(0.019)
post	-0.021	0.079**	-0.017	-0.048***
	(0.031)	(0.037)	(0.012)	(0.014)
PAE*post	0.053	0.091*	-0.005	-0.022
	(0.054)	(0.053)	(0.028)	(0.028)
Controls	No	Yes	No	Yes
Observations	6,558	6,491	6,558	6,491
Panel B: OLS and IPWE				
PAE (OLS)	0.064	0.071	-0.026	-0.026
	(0.042)	(0.044)	(0.023)	(0.023)
Controls	No	Yes	No	Yes
Observations	3,307	3,275	3,307	3,275
PAE (IPWE)	0.090*	0.088*	-0.031	-0.033
	(0.048)	(0.046)	(0.023)	(0.023)
Controls	No	Yes	No	Yes
Observations	2,803	2,795	2,803	2,795
Panel C: Placebo - year 2009				
PAE	-0.003	0.006	-0.023	-0.025
	(0.042)	(0.041)	(0.020)	(0.020)
Controls	No	Yes	No	Yes
Observations	2,721	2,708	2,721	2,708

Note: Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

(1) and (3) and columns (2) and (4), respectively). Overall, they show that, even with this reduced sample, the program had a statistically significant impact on students' rate of decline: the estimated increase on rate of decline in test performance is about 0.09 of one standard deviation. The larger size of the impact of the program is related to differences between the sample used here and that of our main analysis. The school sample reduces by 65% since there are only 137 schools that participated in both PISA 2009 and PISA 2012. We also checked whether the higher impacts are driven by the sample reduction by replicating the main results after imposing difference-in-difference sample restrictions. The magnitude of the coefficients with the smallest sample is the same as in the difference-in-difference approach. The results are reported in Panel B.

The difference-in-difference approach relies on the parallel trends assumption: in the absence of the program, treatment and control schools would have had a parallel trend in the average outcomes of interest. We have pre-treatment information on the rate of decline for the year 2009. We can therefore run a placebo test by comparing treated and control schools in 2009 before the introduction of the PAE, both with an OLS regression and by using the inverse probability weighting estimator. We can test if the outcomes of the two groups of schools were different: significant coefficients in placebo regressions would invalidate this estimation strategy and would question the adequacy of our comparison group. Placebo tests are reported in Panel C of Table 6. We do not find statistically significant coefficients. This can be interpreted as evidence in favor of self-selection being not a real issue in our dataset. If the impact of the program was capturing unobserved teachers' characteristics in those schools that join the program, it should affect students in the pre- and post- treatment periods (assuming that school teams are stable enough). There exists the possibility that school teams improve their characteristics through time, in which case the difference-in-difference analysis could not completely tackle the positive bias. Nevertheless, most literature in teacher value added find that teachers mostly improve their performance in the first two years of work (see Rivkin et al. (2005)).

7 The impact of the program by gender

The second goal of the paper is to analyse the possible existence of heterogeneous effects of remedial programs depending on student's gender. Results for the whole sample can be found in Panel A of Table 7. Columns (1) to (4) show the estimated impact of the program on boys and columns (5) to (8) on girls.

Table 7: The impact of PAE on the rate of decline by gender

	Boys				Girls			
	OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: All sample								
				Level				
PAE	-0.046	-0.022	-0.012	-0.012	0.090***	0.091***	0.106***	0.094***
	(0.030)	(0.029)	(0.032)	(0.031)	(0.028)	(0.028)	(0.032)	(0.030)
Mean in control[a]	-0.046	-0.046	-0.08	-0.08	0.03	0.03	0.01	0.01
				P25 of the entire sample				
PAE	0.020	0.013	0.006	0.006	-0.041***	-0.043***	-0.046***	-0.044***
	(0.013)	(0.013)	(0.014)	(0.014)	(0.011)	(0.012)	(0.013)	(0.013)
Mean in control[a]	0.269	0.269	0.284	0.284	0.238	0.238	0.245	0.245
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,089	10,964	11,011	10,964	11,089	10,964	11,011	10,964
Panel B: Number of treated students in PAE schools								
Fewer Students than the median								
				Level				
PAE	-0.059	-0.039	-0.033	-0.027	0.093***	0.094**	0.098***	0.085**
	(0.044)	(0.039)	(0.046)	(0.040)	(0.036)	(0.037)	(0.037)	(0.038)
				P25 of the entire sample				
PAE	0.020	0.014	0.007	0.003	-0.046***	-0.048***	-0.041***	-0.039***
	(0.019)	(0.018)	(0.020)	(0.019)	(0.014)	(0.015)	(0.014)	(0.015)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	8,986	8,888	8,908	8,888	8,986	8,888	8,908	8,888
More Students than the median								
				Level				
PAE	-0.008	0.021	0.023	0.024	0.152***	0.165***	0.172***	0.170***
	(0.043)	(0.046)	(0.049)	(0.049)	(0.043)	(0.042)	(0.049)	(0.045)
				P25 of the entire sample				
PAE	-0.006	-0.014	-0.019	-0.019	-0.055***	-0.061***	-0.058***	-0.060***
	(0.021)	(0.022)	(0.021)	(0.022)	(0.020)	(0.020)	(0.022)	(0.021)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	8,519	8,425	8,441	8,425	8,519	8,425	8,441	8,425
Test of equality of the coefficients								
Level	no	no	no	no	no	no	reject	reject
P25 of the entire sample	no	no	no	no	no	no	reject	reject

Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

[a] In columns (1), (2), (5) and (6) we report the mean of the control group. In columns (3), (4), (7) and (8) and we report the mean of the weighted control group.

The rate of decline increases by 0.94 of one standard deviation, only for girls. We

find no statistically significant effect among boys. That is, girls that participated in the program experience a lower decline in performance than their similar counterparts who did not participated in it. The program participation also reduced the probability of belonging to the bottom quartile by 4.4 p.p., only among girls.

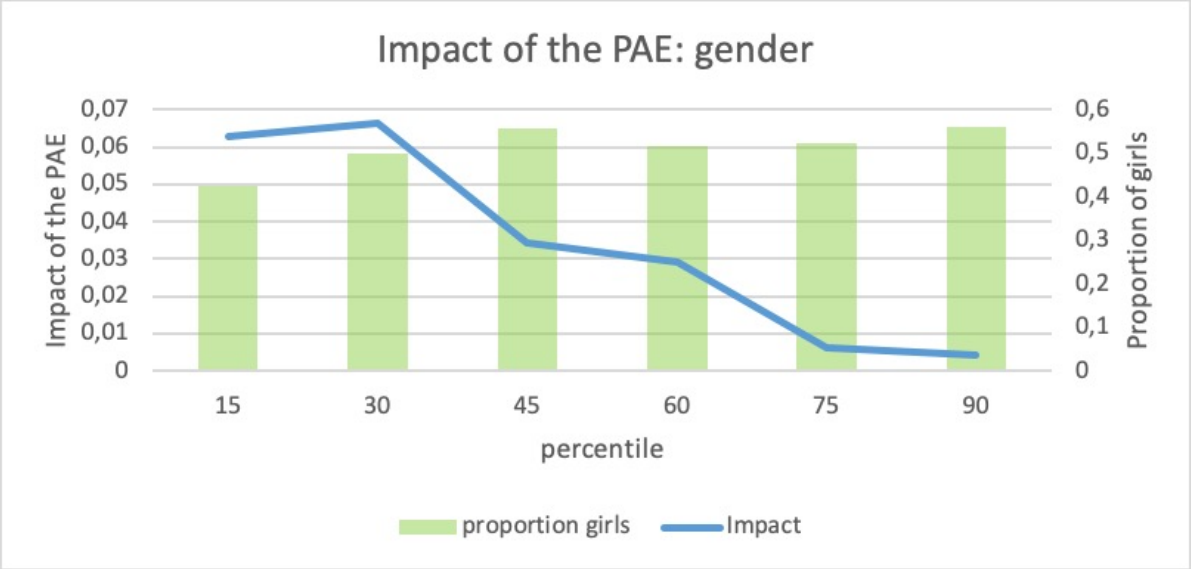
In line with the general analysis, we then investigate whether the number of treated students has a differential effects for boys and girls. Panel B of Table 7 summarizes the main findings. The rate of decline and the probability of falling behind into the bottom part of the distribution for girls are more affected at schools with a larger number of treated students. No statistically significant impact is observed for boys regardless the number of treated students at schools. Overall, we can conclude that the PAE had a substantial positive effect on girls' non-cognitive skills, while boys may also had improved their abilities to sustain test performance but not enough to be identified under the definition of treatment effect at the school level.[34]

We are therefore interested in investigating the potential mechanisms explaining the impact of the program mostly on girls. First, girls could be over-represented in those percentiles in the test performance distribution where the impact of the PAE is larger. In order to check that, we estimate the impact of the PAE along certain percentiles of the rate of decline and the proportion of girls in these same percentiles. To compute the former we calculate the values of two Cumulative Distribution Function (CDF) of the rate decline for certain percentiles: the CDF of rate decline among students in treated schoolsand the CDF of rate decline among re-weighted controls. Next, we present the difference between these two CDF (in particular the absolute value of the rate equal to the CDF treated/CDF weighted controls minus one). Figure 3 shows the results.

The x-axis reports the percentile in the rate decline, while on the y-axes we have both the proportion of girls (histograms) and the impact of the PAE (plot). We observe that the group of students who are more affected is in the lowest tail of the distribution, precisely

[34]The test of equality of the coefficients reported at the end of Panel B indicates whether the estimates for fewer and more students than the median are statistically significantly different. It shows that there are no differences between the coefficients for boys, nor for girls when we use OLS techniques. However, if we use IPWE, as reported in Tables (7) and (8), the coefficients for fewer and more students are statistically significantly different. The null hypothesis of the equality of the coefficients is rejected.

Figure 3: Impact of the PAE: Gender



Note: The x-axis reports the percentile in the rate decline. The y-axes report both the proportion of girls (green histograms) and the impact of the PAE (blue plot).

the students whose rate of decline is lower than the 30 percentile in the distribution. Among these, and also along the entire distribution, girls and boys are evenly distributed. Therefore, the impact of the program only on girls is clearly not due to a larger proportion of girls in the percentiles where its impact is larger.

Second, girls could participate to the remedial program more than boys. The lack of data on individual participation to the program does not allow us to unquestionably exclude this possibility. However, based on observables, this concern is unlikely to apply. The students' characteristics by gender in treated schools are reported in Table 8.

Girls are less likely than boys to show characteristics associated to students targeted by a remedial education intervention. They are less likely to have repeated one or more grades and report a higher index of education possession. If we were expecting a differential participation to the program by gender, boys could participate more than girls.[35]

Third, participation to the program could have lead to gender differences in test taking strategies, where test taking strategies are defined as any strategy that lead students to answer the questions in a different order than the one proposed. However, by using

[35]Recent data regarding individual participation by gender in a very similar remedial program implemented after 2012 in the Region of Madrid reveal that, if any, boys are more likely to participate in these programs than girls.

Table 8: Summary Statistics

Variable	(1) Girls	(2) Boys	(3) P-value Diff. (1)-(2)	(4) P-score Controls
Individual variables				
Initial test score[a]	.588	.625	.000	yes
	(.276)	(.27)		
Migrant(=1)	.152	.146	.606	yes
	(.36)	(.354)		
Repeated once(=1)	.241	.301	.000	yes
	(.428)	(.459)		
Repeated more than once(=1)	.084	.128	.000	yes
	(.278)	(.334)		
Attended kindergarden(=1)	.844	.815	.019	yes
	(.363)	(.388)		
Socioeconomic variables				
Index of education possession[b]	.097	-.014	.000	yes
	(.863)	(.907)		
Mother highly educated(=1)[c]	.3	.307	.653	yes
	(.458)	(.461)		
Father highly educated(=1)[d]	.278	.319	.007	no
	(.448)	(.466)		
School variables				
School size (no. students)	622.138	621.489	.943	yes
	(274.93)	(281.311)		
Prop. of dropout	.114	.115	.832	yes
	(.110)	(.112)		
Prob. of dropout in	.32	.296	.121	yes
high quartile(=1)	(.466)	(.457)		
ESCS[e]	-.375	-.366	.784	no
	(.976)	(.967)		
ESCS in high quartile(=1)	.152	.16	.504	yes
	(.36)	(.367)		
Student-Teacher Ratio	9.259	9.268	.884	yes
	(2.019)	(2.085)		
Parental pressure on teachers(=1)[f]	.401	.382	.236	yes
	(.49)	(.486)		
School climate-teacher(=1)[g]	.696	.677	.221	yes
	(.46)	(.468)		
Rural(=1)[h]	.416	.406	.563	no
	(.493)	(.491)		
Observations	1,843	1,851		

Standard deviations in parentheses.

[a] Initial test score corresponds to the average score in the first five questions of the first cluster of the test.

[b] The index of education possession indicates whether the home possesses a desk and a quiet place to study, a computer and/or educational software and books to help with school work, and a dictionary. It ranges between -3.93 and 1.12.

[c] The mother is defined as highly educated if she has achieved at least tertiary education.

[d] The father is defined as highly educated if he has achieved at least tertiary education.

[e] Index of economic, social, and cultural status.

[f] The dummy is equal to 1 if the principal claims that parents exert pressure into teachers and principal to improve the school quality.

[g] It is a dummy equal to 1 if the school is below the median value of the index of teacher-related factors affecting school climate. Positive values indicate that the teacher-related behaviors hinder learning to a lesser extent. The index ranges between -3.2778 + 2.8533.

[h] It is a dummy equal to 1 if the school is located in a village or a small town.

data also from PISA 2015 (whose test were given on the computer and navigation across question units was restricted), Balart and Oosterveen (2019) disregard the possibility that test taking strategies are a determinant for the gender differences in performance during the test.

We are then left with the last plausible explanation: the possible existence of gender gap in non-cognitive skills prior to the remedial intervention combined with a more effective participation among girls might be behind our result. The gender gap in non-cognitive skills, which is well documented in the literature, could be larger in our setting since disadvantaged backgrounds prevail among students in treated schools and recent evidence shows that the non-cognitive development of boys appears more dependent to parental inputs than girls' (see Bertrand and Pan (2013) among others). If, in addition to that, girls better respond to the PAE then, we can conclude that the remedial education program is more effective in improving non-cognitive skills for them.

The lack of accurate data on teachers' characteristics does not allow to investigate further the mechanisms at play. Nonetheless, recent data on a very similar remedial program implemented after 2012 in the Region of Madrid reveal that, on average, there are twice as female teachers giving remedial education classes than male. If having a female professor acts as a role model for girls and motivate them more, this could partially help to understand our results.

8 Concluding remarks

Recent evidence pointing towards a worsening of the education level of the workforce have called the attention of policy makers and impelled them to improve it (see Carcillo et al. (2015) for an overview of the situation of the most disadvantaged youth in OECD countries). Poor-achieving students are more likely to be early school leavers', which has long-run negative effects, increasing the risk of social exclusion and poverty. National governments are being encouraged to undertake evidence-based education policies to reduce the adverse effects of the aforementioned facts, in line with one of the EU's education

targets for 2020 of reducing the rates of young people leaving early the education and training systems. In this paper, we provide new evidence on these type of interventions by taking advantage of a remedial program aimed at teenagers and recently implemented in Spain, the Program for School Guidance, which offered additional instruction time for underperforming students from poor socioeconomic backgrounds. We tackle the fact that schools' participation in the program cannot completely be considered a random event by using a matching procedure weighting method and a difference-in-difference strategy.

Our main finding is that this program had a substantial positive effect on students' ability to sustain test performance. In particular, it helped girls in improving their rate of decline in performance during the PISA test. It reduced the probability of falling behind into the bottom of the rate of decline distribution by 4.4 p.p. and reduced the decline in performance during the test by almost 0.1 of one standard deviation. As treated schools in our sample participate in the PAE, on average, for three years, should the impact be the same for every year, then the impact of being treated one year could be of 1.47 p.p. and 0.03 of one standard deviation, respectively. We find no statistically significant impact of the program among boys. Therefore, as it is known that improvements in non-cognitive skills have similar effects to cognitive ones for a variety of long-term outcomes (such as job market prospects or higher education investments), the program proved to have a substantially positive impact on the treated youths' life outcomes. This project contributes to the relatively scarce literature on the evaluation of remedial education programs for teenage students on pupils' non-cognitive skills in developed countries. By improving our understanding of the overall effectiveness of remedial education programs, our study might be highly relevant from a policy perspective. It provides a more comprehensive analysis of the strength of such programs.

References

Abadie, A. and G. Imbens (2002). Simple and Bias-Corrected Matching Estimators for Average Treatment Effects. *NBER WP 283*.

Almlund, M., A. Duckworth, J. Heckman, and T. Kautz (2011). Personality Psychology and Economics. In *Handbook of the Economics of Education*, Volume 4 (5), Chapter 1, pp. 1–181. Elsevier.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy 113*(1), 151–184.

Balart, P. and M. Oosterveen (2019). Females show more sustained performance during test-taking than males. *Nature Communications 10, Article number: 3798*.

Balart, P., M. Oosterveen, and D. Webbink (2018). Test Scores, Noncognitive Skills and Economic Growth. *Economics of Education Review 63*, 134–153.

Battaglia, M. and L. Lebedinski (2015). Equal Access to Education: An Evaluation of the Roma Teaching Assistant Program in Serbia. *World Development 76*, 62–81.

Bernstein, D., L. Penner, A. Clarke-Sterwart, and E. Roy (2007). *Psychology*. 8th edition, Wadsworth Publishing.

Bertrand, M. and J. Pan (2013). The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior. *American Economic Journal: Applied Economics 5*(1), 32–64.

Bettinger, E. and B. Long (2009). Addressing the Needs of Under-prepared College Students: Does College Remediation Work? *Journal of Human Resources 44*, 736–771.

Borghans, L., A. Duckworth, J. Heckman, and B. T. Weel (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources 43 (4)*, 972 –1059.

Borghans, L. and T. Schils (2018). Decomposing achievement test scores into measures of cognitive and noncognitive skills. Working Paper Available at SSRN:https://ssrn.com/abstract=3414156.

Brunello, G., A. Crema, and L. Rocco (2018). Testing at Length If It is Cognitive of Non-Cognitive. *IZA DP 11603*.

Busso, M., J. D. Nardo, and J. McCrary (2014). New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators. *Review of Economics and Statistics 96(5)*.

Calcagno, J. and B. T. Long (2008). The Impact of Postsecondary Remediation Using a Regression Discontinuity Approach: Addressing Endogenous Sorting and Noncompliance. *NBER WP 14194*.

Carcillo, S., R. Fernández, S.Königs, and A. Minea (2015). NEET Youth in the Aftermath of the Crisis: Challenges and Policies. Technical report, OECD Social, Employment and Migration Working Papers, No. 164, OECD Publishing Paris.

Carneiro, P. and J. Heckman (2003). Human Capital Policy. *IZA DP 821*.

Cornwell, C., D. B. Mustard, and J. V. Parys (2013). Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School. *Journal of Human Resources 48 (1)*, 236–264.

Cunha, F. and J. Heckman (2008). Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Journal of Human Resources 43(4)*, 738–82.

De Paola, M. and V. Scoppa (2014). The Effectiveness of Remedial Courses in Italy: A Fuzzy Regression Discontinuity Design. *Journal of Population Economics 27(2)*, 365–386.

De Paola, M. and V. Scoppa (2015). Procrastination, Academic Success and the Effectiveness of a Remedial Program. *Journal of Economic Behavior and Organization 115*, 217–236.

Fryer, R. G. and S. D. Levitt (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics 2(2)*, 210–240.

García-Pérez, J. and M. Hidalgo-Hidalgo (2017). No Student Left Behind? Evidence from the Programme for School Guidance in Spain. *Economics of Education Review 60*, 97–111.

Heckman, J. (2000). Policies to Foster Human Capital. *Reseach in Economics 54*, 3–56.

Heckman, J. and Y. Rubinstein (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review 91(2)*, 145–49.

Heckman, J., J. Stixrud, and S. Urzua (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics 24(3)*, 411–82.

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score. *Econometrica 71(4)*, 1161–1189.

Hitt, C., J. Trivitt, and A. Cheng (2016). When You say Nothing at All: The Predictive Power of Student Effort on Surveys. *Economics of Education Review 52*, 105–119.

Holmlund, H. and O. Silva (2014). Targeting Noncognitive Skills to Improve Cognitive Outcomes: Evidence from a Remedial Education Intervention. *Journal of Human Capital 8(2)*, 126–160.

Hospido, L., E. Villanueva, and G. Zamarro (2015). Finance for All: the Impact of Financial Literacy Training in Compulsory Secondary Education in Spain. *Banco de España WP 1502*.

Jacob, B. A. (2002). Where the Boys aren't: Non-cognitive Skills, Returns to School and the Gender Gap in Higher Education. *Economics of Education Review 21(6)*, 589–598.

Lavy, V., A. Kott, and G. Rachkovski (2020). Does Remedial Education at Late Childhood pay off after all? Long-run Consequences for University Schooling, Labor Market Outcomes and Inter-generational Mobility. *Journal of Labor Economics, forthcoming*.

Lavy, V. and A. Schlosser (2005). Targeted Remedial Education for Underperforming Teenagers: Costs and Benefits. *Journal of Labor Economics 23(4)*, 839–874.

Lindqvist, E. and R. Westman (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics 3(1)*, 101–28.

Lord, F. M. (1952). The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties. *Psychometrika 18*, 181–194.

Martins, P. (2017). (How) Do Non-Cognitive Skills Programs Improve Adolescent School Achievement? Experimental Evidence. *CGE Working Paper 81*.

Nollenberger, N., N. Rodríguez-Planas, and A. Sevilla (2016). The Math Gender Gap: The Role of Culture. *American Economic Review 106(5)*, 257–261.

OECD (2013). Crisis Squeezes Income and puts Pressure on Inequality and Poverty. Technical report, OECD Publishing, Paris.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics 37*(2), 187–204.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic Achievement. *Econometrica 73(2)*, 417–458.

Rodríguez-Planas, N. and N. Nollenberger (2018). Let the Girls Learn! It is not Only about Math.. It's about Gender Social Norms. *Economics of Education Review 62*, 230–153.

Sternberg, R. J., G. Forsythe, J. A. Hedlund, R. Horvath, R. K. Wagner, W. Williams, S. A. Snook, and E. Grigorenko (2000). Practical Intelligence in Everyday Life. New York, NY, Cambridge University Press.

ter Weel, B. (2008). The Noncognitive Determinants of Labor Market and Behavioral Outcomes: Introduction to the Symposium. *Journal of Human Resources 43(4)*, 729–37.

Zamarro, G., A. Cheng, M. D. Shakeel, and C. Hitt (2018). Comparing and Validating Measures of Non-Cognitive Traits: Performance Task Measures and Self-reports from a Nationally Representative Internet Panel. *Journal of Behavioral and Experimental Economics 72(C)*, 51–60.

Zamarro, G., C. Hitt, and I. Mendez (2019). When Students Don't Care: Reexamining International Differences in Achievement and Student Effort. *Journal of Human Capital 13*(4), 519–552.

Online Supplementary Material to:

Non-Cognitive Skills and Remedial Education: Good News for Girls

A: Self-assessed measures

We examine here the impact of the program on several students' self-assessed measures. First, we consider absenteeism and truancy, defined as whether the student does not show up at school or is usually late for it. This information is relevant since it is likely correlated with motivation and it may also predict worse test scores. The more one misses classes, the less likely can be motivated to learn or find it more difficult. Second, discipline is measured by the way students behave in class (disciplinary climate). Third, self-confidence is measured by self-reported ability to succeed with enough effort and confidence to perform well if wanted. Another way to measure self-confidence is sense of belonging to the group, in our case the school. We finally look at motivation towards schools: whether students think that school does prepare for life or it is considered a waste of time, and if it helps to get a job and improve career chances. Summary statistics for these variables can be found in Table A.1 below.

Overall, we observe that absenteism reduces and discipline improves, the latter especially for boys. Also, the perception of learning at school increases. The rest of the measures are not precisely estimated, although the sign of the coefficients are as expected (Table A.2). Recall that these measures are self-assessed and we cannot know a priori whether the exposure to the program affected differently their perception or simply what they do report. However, once more results are positive and in line with findings for similar interventions in the US as reviewed by Heckman (2000).

Table A.1: Students' outcomes: self-assessed non-cognitive skills

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Boys				Girls				Overall			
	All	Treated	Control	Weighted Control	All	Treated	Control	Weighted Control	All	Treated	Control	Weighted Control
Motivation												
Absenteism(=1)	.238	.249	.232	.274	.241	.253	.235	.281	.239	.251	.233	.277
	(.426)	(.433)	(.422)	(.446)	(.428)	(.435)	(.424)	(.45)	(.427)	(.434)	(.423)	(.448)
Observations	5,406	1,823	3,574	3,574	5,567	1,832	3,735	3,735	10,973	3,664	7,309	7,309
Truancy(=1)	.366	.385	.357	.381	.364	.382	.355	.385	.365	.383	.356	.383
	(.482)	(.487)	(.479)	(.486)	(.481)	(.486)	(.479)	(.487)	(.481)	(.486)	(.479)	(.486)
Observations	5,374	1,823	3,551	3,551	5,537	1,818	3,719	3,719	10,911	3,641	7,270	7,270
Discipline												
Bad climate(=1)	.39	.372	.399	.41	.357	.343	.364	.366	.373	.358	.381	.388
	(.488)	(.484)	(.49)	(.492)	(.479)	(.475)	(.481)	(.482)	(.484)	(.479)	(.486)	(.487)
Observations	5,441	1,851	3,590	3,590	5,584	1,843	3,741	3,741	11,025	3,694	7,331	7,331
Self-confidence(=1)	.287	.279	.291	.292	.26	.251	.265	.26	.273	.265	.278	.276
	(.452)	(.449)	(.454)	(.455)	(.439)	(.434)	(.441)	(.438)	(.446)	(.441)	(.448)	(.447)
Observations	5,441	1,851	3,590	3,590	5,584	1,843	3,741	3,741	11,025	3,694	7,331	7,331
Sense of belonging(=1)	.952	.956	.95	.951	.973	.974	.973	.967	.963	.965	.962	.959
	(.214)	(.205)	(.219)	(.217)	(.162)	(.159)	(.163)	(.178)	(.188)	(.183)	(.191)	(.198)
Observations	2,990	1,021	1,969	1,969	3,382	1,116	2,266	2,266	6,372	2,137	4,235	4,235
Perception of learning	.587	.586	.587	.584	.641	.647	.638	.633	.614	.616	.631	.608
at school(=1)	(.492)	(.493)	(.492)	(.493)	(.48)	(.478)	(.481)	(.482)	(.487)	(.486)	(.487)	(.488)
Observations	5,441	1,851	3,590	3,590	5,584	1,843	3,741	3,741	11,025	3,694	7,331	7,331

Note: Standard deviations in parentheses.

Table A.2: The impact of PAE on self-assessed non-cognitive outcomes

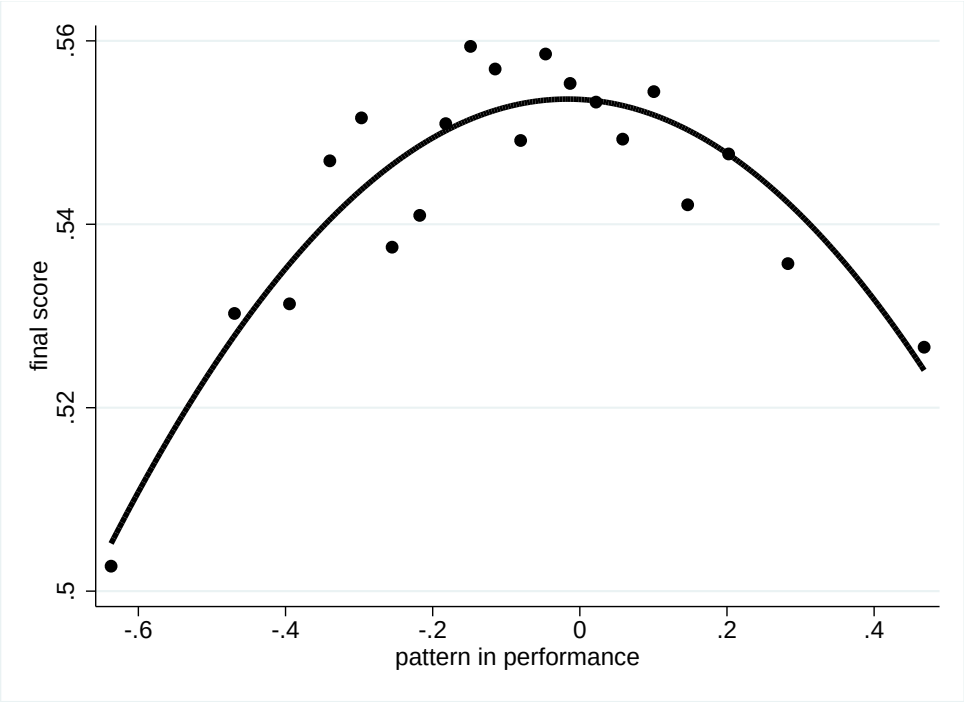
	Boys				Girls				Overall			
	OLS		IPWE		OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Motivation												
						Absenteism						
PAE	0.016	-0.020	-0.024	-0.015	0.018	-0.023	-0.028	-0.031*	0.017	-0.022	-0.026	-0.025*
	(0.017)	(0.015)	(0.019)	(0.018)	(0.019)	(0.016)	(0.02)	(0.017)	(0.015)	(0.014)	(0.017)	(0.015)
						Truancy						
PAE	0.026	0.007	0.003	0.009	0.026	0.004	-0.003	-0.002	0.026	0.006	0.002	0.004
	(0.017)	(0.017)	(0.019)	(0.018)	(0.020)	(0.019)	(0.022)	(0.020)	(0.015)	(0.017)	(0.017)	(0.015)
Discipline												
						Bad Climate						
PAE	-0.025*	-0.030*	-0.037**	-0.035**	-0.018	-0.025	-0.022	-0.021	-0.021*	-0.027**	-0.030**	-0.028**
	(0.014)	(0.015)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)	(0.016)	(0.011)	(0.012)	(0.012)	(0.012)
						Self-confidence						
PAE	-0.010	-0.010	-0.013	-0.015	-0.014	-0.011	-0.009	-0.008	-0.012	-0.011	-0.011	-0.011
	(0.012)	(0.012)	(0.013)	(0.013)	(0.012)	(0.013)	(0.014)	(0.014)	(0.009)	(0.009)	(0.010)	(0.009)
						Sense of belonging						
PAE	0.006	0.008	0.005	0.004	0.002	0.005	0.007	0.008	0.004	0.008	0.008	0.007
	(0.008)	(0.009)	(0.009)	(0.009)	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)
						Perception of learning at school						
PAE	0.000	0.001	0.001	0.008	0.013	0.016	0.014	0.016	0.006	0.012*	0.008	0.012*
	(0.013)	(0.013)	(0.014)	(0.014)	(0.012)	(0.012)	(0.013)	(0.013)	(0.007)	(0.006)	(0.008)	(0.007)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,025	10,941	10,973	10,941	11,025	10,941	10,973	10,941	11,025	10,941	10,973	10,941

Note: Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

B: Cognitive Skills

We present first the relationship between students' final scores in the test and their pattern in performance through it:[36]

Figure B.1: Pattern of performance and final score



Note: The figure represents a binned scatter plot of the relationship between final score and individual pattern of performance.

As it can be observed, there is a non-lineal relationship between final score and pattern in performance. Final scores are low both for students with very negative and very positive pattern of performance. As the pattern of performance improves, final score increases but only up to some value. Further improvements in performance imply reductions in the final score. The reason might be that those students with a positive rate of decline are those with worse initial score (as mentioned in the main text). As the test moves on, they perform better but this improvement does not compensate the bad initial score. Therefore, in order to analyze the impact of the program on students' rate of decline (or

[36]We use the term final score to refer to the average number of correct answers in the PISA test and not to the actual scores provided by PISA. PISA uses weights based on cognitive response theory. In particular, it uses cognitive item theory and provide several plausible values for each of the competences being evaluated (see OECD, 2012). It is therefore not possible to establish a direct relationship between average number of correct answers in the PISA test and the actual PISA test score. Nevertheless, the correlation between the average number of correct answers and the PISA measures is high, in particular larger than 0.8 and statistically significant at 1%.

score) in isolation it is crucial to control for their individual score (or, respectively, rate of decline), as we do in this paper. We present here results on the impact of the PAE on initial and final score while controlling for individual rate of decline.

Table B.1: The impact of PAE on initial and final performance

	OLS		IPWE			NNPS		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Initial Performance				
				Level				
PAE	-0.078***	0.026	0.031	0.027	0.073***	0.064***	0.057**	0.053**
	(0.028)	(0.022)	(0.030)	(0.024)	(0.025)	(0.023)	(0.023)	(0.022)
				P25 of the entire sample				
PAE	0.028***	-0.001	-0.000	0.001	-0.013	-0.011	-0.008	-0.005
	(0.009)	(0.008)	(0.011)	(0.009)	(0.010)	(0.009)	(0.009)	(0.009)
Observations	11,038	10,964	10,991	10,964	10,964	10,964	10,964	10,964
				Final Performance				
				Level				
PAE	-0.102***	0.022	0.040	0.031	0.040	0.041*	0.052**	0.056***
	(0.036)	(0.027)	(0.039)	(0.028)	(0.024)	(0.022)	(0.022)	(0.021)
				P25 of the entire sample				
PAE	0.032**	-0.011	-0.015	-0.012	-0.007	-0.008	-0.014	-0.015
	(0.014)	(0.011)	(0.015)	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	Yes	Yes	Yes	Yes
No. matches per obs.	-	-	-	-	2	4	6	8
Observations	11,051	10,977	11,004	10,977	10,977	10,977	10,977	10,977

Note: Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

In addition to the IPWE, we also compare each treated student with her most similar associated control counterparts and thus provide results using several nearest neighbor propensity score estimators. In particular, we provide estimators by varying the number of nearest neighbors considered in the estimation from 2 to 8 (NNPS(2) to NNPS(8) in columns 5 to 8). Note that, the results using NNPS are similar to those obtained by using the inverse probability weighting estimator. As it can be observed, the program increases students' initial score on average (however it does not improve the initial score among those in the bottom part of the distribution). Therefore, as the program improves student's rate of decline on average, it increases students' final score on average in line with previous results in the literature (García-Pérez and Hidalgo-Hidalgo, 2017).

C: Alternative definitions of difficulty and order of the subjects

We use an alternative definition of difficulty of the question: the percentage of students who correctly answer the question. The results are consistent with those in the main text and are reported in Table C.1.

Table C.1: The impact of PAE on the rate of decline - Different measure of difficulty

	Boys				Girls				Overall			
	OLS		IPWE		OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Difficulty as percentage of students who correctly answer to the question												
	Level											
PAE	-0.057*	-0.036	-0.020	-0.023	0.089***	0.095***	0.112***	0.101***	0.015	0.030	0.044*	0.039*
	(0.031)	(0.031)	(0.034)	(0.034)	(0.027)	(0.027)	(0.030)	(0.029)	(0.022)	(0.023)	(0.024)	(0.023)
	P25 of the entire sample											
PAE	0.028*	0.019	0.008	0.010	-0.044***	-0.050***	-0.053***	-0.053***	-0.008	-0.016*	-0.022**	-0.021**
	(0.014)	(0.014)	(0.015)	(0.015)	(0.011)	(0.012)	(0.013)	(0.013)	(0.009)	(0.009)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,091	10,963	11,012	10,963	11,091	10,963	11,012	10,963	11,091	10,963	11,012	10,963

Note: Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

We also recode a question as correct if the answer is correct or partially correct. The results are again consistent and available upon request.

In addition, we analyze whether, within a single booklet, question position is correlated with question characteristics. For instance, students assigned booklet 3, will all have reading questions only at the beginning of the test which might imply a position correlation between position and the subject. We check here whether the order of the subjects, that is, whether maths is taken before reading and vice versa, could be relevant for differences in the rate of decline. As reported in Table C.2 the order of the subject does not show to be relevant for the rate of decline. We observe that, independently of the order of clusters, the remedial program benefits slightly more girls than boys and that by gender taking reading after maths or viceversa is not statistically relevant (p-value of Chi2 test for equality in coefficients).

Finally, as an additional check, we also compute the pattern of performance by estimating equation (1) for each school (instead of each student). This allows us to add

Table C.2: The impact of PAE on the rate of decline - Clusters order

	Boys		Girls		Overall		
	OLS	IPWE	OLS	IPWE	OLS	IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	Observations
				Level			
Reading after Maths	-0.015	-0.004	0.114**	0.102**	0.050	0.048	4,238
	(0.048)	(0.049)	(0.046)	(0.049)	(0.033)	(0.031)	
Maths after Reading	-0.027	-0.019	0.066**	0.077**	0.021	0.031	6,726
	(0.036)	(0.038)	(0.033)	(0.037)	(0.025)	(0.024)	
Chi2 test (p-value)	0.1832	0.0074	0.5243	0.1251			
				P25 of the entire sample			
Reading after Maths	0.026	0.015	-0.046**	-0.038*	-0.009	-0.010	4,238
	(0.021)	(0.021)	(0.020)	(0.020)	(0.015)	(0.014)	
Maths after Reading	0.007	-0.001	-0.042***	-0.047***	-0.018	-0.025**	6,726
	(0.016)	(0.016)	(0.015)	(0.017)	(0.011)	(0.011)	
Chi2 test (p-value)	0.5876	0.0180	0.9998	0.8653			

Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

booklet-specific fixed effects. By doing so, we can isolate the pattern of performance while netting out variation due to ordering in terms of difficulty. However, notice that this approach does not allow us to compute individual pattern of performance. We thus compare our measure of pattern of performance in the main text with the one obtained including booklet-specific fixed effects in Table C.3. As it can be observed, they are no statistically different for the all sample nor control or treated students.

Table C.3: Rate of performance and difficulty

	All	Treated	Control
	(1)	(2)	(3)
Mean in main text	-.096	-.092	-.098
	(.264)	(.266)	(.264)
Observations	11,011	3,684	7,327
Mean with booklet and school fixed effects	-.098	-.094	-.099
	(.051)	(.051)	(.051)
Observations	11,025	3,694	7,331
P-value of the difference between the coefficients	0.5176	0.6231	0.6560

Standard deviations in parentheses.

D: Item reached

We consider another non-self assessed measure: the number of items reached in the test which corresponds to the average last question answered by the student in each of the four clusters. Table D.1 provides the summary statistics for the average number of items reached and Table D.2 the impact of the PAE on this outcome.

Table D.1: Students' outcomes: item reached

	Boys				Girls				Overall			
	All	Treated	Control	Weighted Control	All	Treated	Control	Weighted Control	All	Treated	Control	Weighted Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Level	.973	.971	.973	.967	.974	.97	.974	.971	.973	.97	.975	.969
	(.079)	(.079)	(.068)	(.077)	(.067)	(.073)	(.065)	(.063)	(.068)	(.076)	(.063)	(.071)
First quartile (P25)	.285	.290	.283	.319	.297	.307	.293	.328	.292	.299	.288	.324
Observations	5,430	1,843	3,587	3,587	5,581	1,841	3,771	3,771	11,011	3,684	7,327	7,327

Note: Standard deviations in parentheses.

As it can be observed, the program has no impact on the number of items reached overall, although it has a slightly statistically significant positive impact for boys. In addition, it reduced the probability of belonging to the bottom quartile in the distribution of item reached by 2.4 p.p. The results are consistent to choosing the minimum or the maximum last question answered. They are not reported but are available upon request.

Table D.2: The impact of PAE on item reached

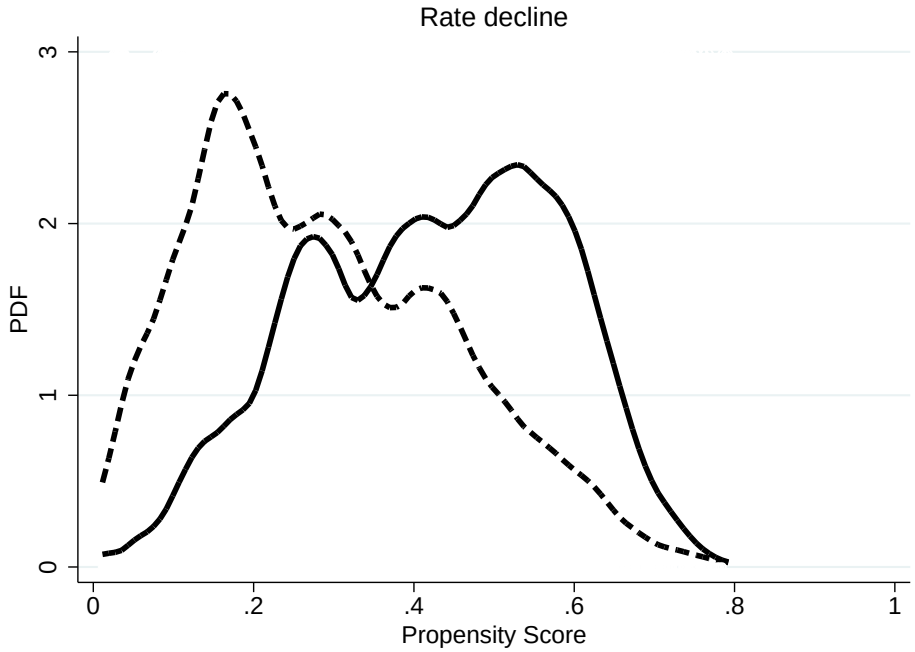
	Boys				Girls				Overall			
	OLS		IPWE		OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Level											
PAE	-0.000	0.004	0.004	0.005*	-0.004*	-0.000	-0.001	0.000	-0.003	0.002	0.001	0.003
	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)
	P25 of the entire sample											
PAE	0.005	-0.021	-0.029	-0.025	0.013	-0.015	-0.021	-0.022	0.009	-0.018	-0.025*	-0.024*
	(0.016)	(0.015)	(0.018)	(0.016)	(0.016)	(0.015)	(0.018)	(0.016)	(0.013)	(0.012)	(0.015)	(0.013)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,089	10,964	11,011	10,964	11,089	10,964	11,011	10,964	11,089	10,964	11,011	10,964

Note: Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

E: Propensity Score Support

Although the two distributions differ in form, the figure shows how similar the control and treatment samples are. The support of the values of the propensity score of stu-

Figure E.1: Propensity Score Support



Note: PDF of the propensity score of treated (solid line) and control students (dotted line).

dents in treated schools (solid line) and that of the control schools (dotted line) are the same: both ranges from 0 to approximately 0.8. In addition, there is no concentration of predicted values around zero or one (which would mean that there are no comparable control students for some treated students).

F: Participation in the remedial program

We estimate the predicted probability of participation in the program as a function of a set of characteristics of the students, parents and schools, i.e., the propensity score, $p(X_i)$. The set of variables included in X_i was chosen according to the differences in mean covariates in Table 2 of the main text. We include the initial test score, measured as the average score in the first five questions of the first cluster, to control for student's cognitive abilities.[37] We also control for gender, immigrant status, whether the student repeated a grade once or for more than one academic year, and whether the student attended pre-primary education. Regarding socioeconomic variables, we include the mother education level and the index of educational materials at home.[38] Finally, we also add a set of school characteristics, including its size, its mean socioeconomic index, the student-teacher ratio, the proportion of dropouts, an indicator of whether teachers favor good school climate and whether parents exert pressure on teachers.[39] The final specification is shown in Table F.1 which presents the estimates of the propensity score for the treatment. Its weights are used to estimate the impact of PAE on the rate of decline.

The mean initial test score at the school level does not affect the probability that the school offer the program, neither it does the proportion of boys in a school. On the contrary, schools with a high percentage of migrants or grade-repeaters are more likely to offer the program than other schools. Observe that, once a complete set of control variables is considered, both parental education and the index of educational materials at home do not seem to influence the probability of being treated. Those schools with poorer socioeconomic index, larger size in terms of number of students, and a larger index of school climate have a higher chance of being treated.

[37] Nonetheless, excluding such variable from the analysis does not change the results.

[38] We also averaged out at school level individual variables, and results do not change. They are available upon request.

[39] We acknowledge that the last two variables could be potentially affected by the policy. We therefore also run the same specification by excluding them from the analysis. The results do not qualitatively change.

Table F.1: Propensity score estimation - Probability of being treated

	(1) Probability of being treated
Individual variables	
Initial test score[a]	0.077
	(0.094)
Girl(=1)	-0.041
	(0.040)
Migrant(=1)	0.455***
	(0.149)
Repeated once(=1)	0.148**
	(0.060)
Repeated more than once(=1)	0.201**
	(0.103)
Attended kindergarden(=1)	-0.045
	(0.098)
Socioeconomic variables	
Index of education possession[b]	0.004
	(0.033)
Mother highly educated(=1)[c]	-0.039
	(0.078)
School variables	
School size (no. students)	0.004**
	(0.002)
Prob. dropouts in high quartile(=1)	0.363
	(0.268)
ESCS[d]	-1.015***
	(0.329)
Student-Teacher Ratio	-0.033
	(0.022)
Parental pressure on teachers(=1)[e]	0.269
	(0.257)
School climate-teacher(=1)[f]	0.599**
	(0.253)
Observations	10,975

Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. We also include regions, interactions between regions and some individual characteristics and school size squared.

[a] Initial test score corresponds to the average score in the first five questions of the first cluster of the test.

[b] The index of education possession indicates whether the home possesses a desk and a quiet place to study, a computer and/or educational software and books to help with school work, and a dictionary. It ranges between -3.93 and 1.12.

[c] The mother is defined as highly educated if she has achieved at least tertiary education.

[d] Index of economic, social, and cultural status.

[e] The dummy is equal to 1 if the principal claims that parents exert pressure into teachers and principal to improve school quality.

[f] It is a dummy equal to 1 if the school is below the median value of the index of teacher-related factors affecting school climate. Positive values indicate that the teacher-related behaviors hinder learning to a lesser extent. The index ranges between -3.2778 + 2.8533.

G: Sub-sample analysis

We focus on two sub-samples of our complete treated students group. We split it according to some pre-treatment characteristics: the proportion of migrants at school and the parental education level. These variables are appropriate as, even though they affect the probability of participating in the PAE, they are not included in the propensity score estimation. This allows us to use the same specification for the propensity score as in the rest of the paper and get comparable results. First, we consider students in treated schools where the proportion of migrants or non-educated parents is above its sample median. Second, we consider students in treated schools where the proportion of migrants or non-educated parents is below its sample median. By considering students in these types of schools, we first increase and then reduce the likelihood that students in our sample attending those schools actually participated in the program. Results are in line with the main findings above: the impact of the PAE is larger in the sub-sample of schools with a larger number of potentially treated students.

Disadvantaged students

First, we consider students in treated schools where the proportion of migrants is above the median value of the distribution of this variable for all public schools. By considering students in these types of schools, we increase the likelihood that they actually participated in the program. Similarly, we consider students in treated schools with non-educated parents.[40] The first four columns of Table G.1 provide results for the impact of the program on the rate of decline and the probability of falling into the bottom quartile of the rate of decline distribution. Rows (2) to (4) provide results for the sub-sample of students at schools with the proportion of migrants above the median. Rows (6) to (8) provide results for the sub-sample of students in non-educated families.

[40]In this analysis only students in treated schools are split into two sub-samples. Alternatively, we could split both treated and controls into two sub-samples. Results of this alternative exercise, available upon request, are similar to the ones found here. This is because control students at schools with a proportion of migrants above the median might not be that similar to students in treated schools and thus receive a low weight. A similar reasoning can be applied to the results found for the sub-sample of students with non-educated parents. Parents are defined as non-educated if their level of education is lower or equal to secondary school.

Table G.1: The impact of PAE on the rate of decline (Subgroups)

	OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	Schools with many migrants				**Schools with few migrants**			
	Level							
PAE	0.013	0.040	0.045	0.041	0.026	0.025	0.043	0.031
	(0.025)	(0.026)	(0.029)	(0.028)	(0.032)	(0.032)	(0.034)	(0.032)
	P25 of the entire sample							
PAE	-0.012	-0.021**	-0.027**	-0.026**	-0.005	-0.006	-0.008	-0.007
	(0.010)	(0.011)	(0.012)	(0.011)	(0.013)	(0.014)	(0.014)	(0.013)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	9,835	9,722	9,757	9,722	8,757	8,665	8,679	8,665
	Non-educated families				**Educated families**			
	Level							
PAE	0.002	0.022	0.034	0.026	0.044	0.049	0.059*	0.052*
	(0.027)	(0.027)	(0.032)	(0.030)	(0.030)	(0.030)	(0.034)	(0.031)
	P25 of the entire sample							
PAE	-0.005	-0.014	-0.019	-0.019	-0.014	-0.015	-0.017	-0.016
	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)	(0.013)	(0.015)	(0.014)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	9,513	9,394	6,745	6,704	9,056	8,952	8,978	8,952

Schools with many (few) migrants are those where the proportion of migrants is above (below) the median value of this variables for all public schools. Non-educated families are those in which the education level of either father or mother is lower or equal to secondary school. Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

Although not precisely estimated, the impact of the program on the rate of decline is an increase of 0.041-0.045 of one standard deviation, in the sub-sample of schools with migrants above the median, which is very close to the impact on the full sample of students. The probability of belonging to the bottom quartile is reduced by between 2.6 and 2.7 p.p overall, when considering schools with migrants above the median. Thus, again, by considering the full sample of students, we came close to estimating the true impact of the PAE on moving students out of low-skills status, which is the main objective of the program. The overall impact of the program is less precisely estimated when considering the sample of students with non-educated families, but confirms the previous results. The coefficients are in line with those obtained with the subsample of schools with migrants above the median and with the full sample, but standard errors are bigger.

Privileged students

Next, we consider students in treated schools where the proportion of migrants is below the median value of the distribution. By considering students in these types of schools we reduce the likelihood that they actually participated in the program. Similarly, we consider students in treated schools with educated parents. The last four columns of Table G.1 provides results for the impact of the program on the rate of decline and the probability of falling into the bottom quartile of the rate decline distribution.

As it can be observed, no impact of the program is found among students in schools with a low proportion of migrants, in line with García-Pérez and Hidalgo-Hidalgo (2017). However, among students with educated parents the impact on the rate of decline is slightly larger than among the whole sample (0.052 vs. 0.041). There are two main explanations to this finding. On the one hand, if the number of true treated students in these schools was indeed low, then we are capturing spillover effects: students who participated have positively benefited the rest. On the other hand, if the number of true treated students in these schools was, as opposed to what we expect, high, then we are capturing the direct effect of the program which seems to be slightly larger among students with educated parents (that is, PAE and parental education are complements).

H: Schools' information on PAE implementation

Table H.1 below compares schools with more and less treated students than the median. Schools with more treated students seem to have a larger number of students from disadvantaged backgrounds than schools with fewer remedial students: they have a larger proportion of migrants, a lower proportion of educated fathers, larger school size, a larger proportion of dropouts and a higher proportion of parents exerting pressure on teachers.

We next compute, for each treated school, the PAE student-teacher ratio as the number of students involved in the remedial program per teacher and/or monitor. On average, there are 9 students per teacher. Table H.2 provides results for the impact of the PAE on our outcomes of interest depending on the school student-teacher ratio in remedial classes. First, we compare students in schools whose PAE student-teacher ratio is higher than its sample median to students in control schools (top panel). Second, we compare students in schools whose PAE student-teacher ratio is lower than its sample median to students in control schools (bottom panel). A low PAE student-teacher ratio should be favorable as it suggests a better quality in the implementation of the program. When comparing performance in both type of schools, in particular on the level of rate of decline, we find that the impact of the program in low and high student-teacher ratio schools are not statistically significant different, suggesting that on average the effect is comparable for the two types of schools (test of equality of the coefficients). However, the benefits on the ability to sustain the test performance are much higher for underperforming students (less than P25 of the entire sample) who are in schools with a low PAE student-teacher ratio.

Table H.1: Summary statistics

Variable	(1) All	(2) Fewer Students than the median	(3) More Students than the median	(4) P-value Diff. (2)-(3)
Individual variables				
Initial test score[a]	.606	.602	.611	.419
	(.276)	(.278)	(.274)	
Girl(=1)	.505	.511	.497	.459
	(.5)	(.5)	(.5)	
Migrant(=1)	.16	.136	.194	.000
	(.367)	(.343)	(.395)	
Repeated once(=1)	.276	.265	.291	.129
	(.447)	(.441)	(.455)	
Repeated more than once(=1)	.108	.119	.092	.024
	(.311)	(.324)	(.29)	
Attended kindergarden(=1)	.819	.81	.831	.156
	(.385)	(.392)	(.375)	
Socioeconomic variables				
Index of education possession[b]	.045	.027	.071	.203
	(.892)	(.892)	(.893)	
Mother highly educated(=1)[c]	.314	.317	.309	.634
	(.464)	(.466)	(.462)	
Father highly educated(=1)[d]	.301	.314	.283	.078
	(.459)	(.464)	(.45)	
School variables				
School size (no. students)	599.179	521.599	709.529	.000
	(261.141)	(230.315)	(262.781)	
Prop. of dropout	.118	.094	.153	.000
	(.113)	(.108)	(.112)	
Prob. of dropout in	.337	.235	.483	.000
high quartile(=1)	(.473)	(.424)	(.5)	
ESCS[e]	-.354	-.377	-.321	.145
	(.978)	(.971)	(.987)	
ESCS in high quartile(=1)	.217	.247	.175	.000
	(.413)	(.431)	(.38)	
Student-Teacher Ratio	9.203	8.731	9.873	.000
	(2.142)	(1.958)	(2.214)	
Parental pressure on teachers(=1)[f]	.402	.354	.471	.000
	(.49)	(.478)	(.499)	
School climate-teacher(=1)[g]	.667	.663	.673	.588
	(.471)	(.473)	(.469)	
Rural(=1)[h]	.449	.552	.303	.000
	(.497)	(.497)	(.46)	
Observations	2,701	1,586	1,115	

Standard deviations in parentheses.

[a] Initial test score corresponds to the average score in the first five questions of the first cluster of the test.

[b] The index of education possession indicates whether the home possesses a desk and a quiet place to study, a computer and/or educational software and books to help with school work, and a dictionary. It ranges between -3.93 and 1.12.

[c] The mother is defined as highly educated if she has achieved at least tertiary education.

[d] The father is defined as highly educated if he has achieved at least tertiary education.

[e] Index of economic, social, and cultural status.

[f] The dummy is equal to 1 if the principal claims that parents exert pressure into teachers and principal to improve the school quality.

[g] It is a dummy equal to 1 if the school is below the median value of the index of teacher-related factors affecting school climate. Positive values indicate that the teacher-related behaviors hinder learning to a lesser extent. The index ranges between -3.2778 + 2.8533.

[h] It is a dummy equal to 1 if the school is located in a village or a small town.

Table H.2: The impact of PAE on the rate of decline (PAE Student-Teacher ratio)

	OLS		IPWE	
	(1)	(2)	(3)	(4)
PAE Student-Teacher ratio lower than the median				
	Level			
PAE	0.035	0.064*	0.070*	0.065*
	(0.035)	(0.037)	(0.039)	(0.037)
	P25 of the entire sample			
PAE	-0.030**	-0.039***	-0.041***	-0.041***
	(0.013)	(0.013)	(0.014)	(0.014)
Controls	No	Yes	No	Yes
Observations	8,739	8,646	8,661	8,646
PAE Student-Teacher ratio higher than the median				
	Level			
PAE	0.041	0.042	0.050*	0.042
	(0.026)	(0.027)	(0.026)	(0.027)
	P25 of the entire sample			
PAE	-0.006	-0.007	-0.009	-0.009
	(0.012)	(0.013)	(0.013)	(0.013)
Controls	No	Yes	No	Yes
Observations	8,821	8,722	8,743	8,722
Test of equality of the coefficients				
Level	no	no	no	no
P25 of the entire sample	no	reject	reject	reject

Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

Table H.3: Summary statistics schools' information on the implementation of the PAE

Variable	(1) With Information	(2) Without Information	(3) P-value Diff. (2)-(1)
Individual variables			
Initial test score[a]	.594	.597	.857
	(.094)	(.08)	
Girl(=1)	.512	.471	.086
	(.113)	(.136)	
Migrant(=1)	.188	.164	.513
	(.212)	(.194)	
Repeated once(=1)	.295	.262	.259
	(.16)	(.152)	
Repeated more than once(=1)	.108	.147	.255
	(.077)	(.219)	
Attended kindergarden(=1)	.805	.857	.030
	(.151)	(.112)	
Socioeconomic variables			
Index of education possession[b]	.006	.062	.314
	(.297)	(.298)	
Mother highly educated(=1)[c]	.301	.262	.132
	(.149)	(.134)	
Father highly educated(=1)[d]	.291	.265	.361
	(.15)	(.152)	
School variables			
School size (no. students)	584.893	631.024	.426
	(274.045)	(327.921)	
Prop. of dropout	.125	.126	.984
	(.117)	(.115)	
Prob. of dropout in	.333	.32	.880
high quartile(=1)	(.474)	(.47)	
ESCS[e]	-.397	-.499	.179
	(.406)	(.408)	
ESCS in high quartile(=1)	.202	.073	.029
	(.404)	(.256)	
Student-Teacher Ratio	9.06	9.292	.535
	(2.257)	(1.859)	
Parental pressure on teachers(=1)[f]	.393	.316	.389
	(.491)	(.468)	
School climate-teacher(=1)[g]	.69	.614	.395
	(.465)	(.493)	
Rural(=1)[h]	.464	.368	.296
	(.502)	(.484)	
Observations	84	44	

Standard deviations in parentheses.

[a] Initial test score corresponds to the average score in the first five questions of the first cluster of the test.

[b] The index of education possession indicates whether the home possesses a desk and a quiet place to study, a computer and/or educational software and books to help with school work, and a dictionary. It ranges between -3.93 and 1.12.

[c] The mother is defined as highly educated if she has achieved at least tertiary education.

[d] The father is defined as highly educated if he has achieved at least tertiary education.

[e] Index of economic, social, and cultural status.

[f] The dummy is equal to 1 if the principal claims that parents exert pressure into teachers and principal to improve the school quality.

[g] It is a dummy equal to 1 if the school is below the median value of the index of teacher-related factors affecting school climate. Positive values indicate that the teacher-related behaviors hinder learning to a lesser extent. The index ranges between -3.2778 + 2.8533.

[h] It is a dummy equal to 1 if the school is located in a village or a small town.

I: School level analysis

Table I.1: Summary statistics at the school level

Variable	(1) All	(2) Treated	(3) Controls	(4) P-value Diff. (2)-(3)	(5) Weighted Controls	(6) P-value Diff. (2)-(4)	(7) P-score
Individual variables							
Initial test score[a]	.608	.594	.615	.041	.614	.314	yes
	(.096)	(.089)	(.099)		(.07)		
Girl(=1)	.501	.497	.503	.604	.501	.820	yes
	(.123)	(.123)	(.123)		(.091)		
Migrant(=1)	.125	.178	.099	.000	.177	.125	yes
	(.17)	(.204)	(.144)		(.201)		
Repeated once(=1)	.247	.284	.229	.001	.261	.392	yes
	(.144)	(.156)	(.134)		(.102)		
Repeated more than once(=1)	.104	.123	.095	.058	.113	.443	yes
	(.138)	(.143)	(.135)		(.094)		
Attended kindergarden(=1)	.834	.825	.839	.345	.823	.590	yes
	(.138)	(.14)	(.137)		(.123)		
Socioeconomic variables							
Index of education possession[b]	.036	.025	.042	.588	.067	.318	yes
	(.3)	(.295)	(.303)		(.228)		
Mother highly educated(=1)[c]	.327	.288	.346	.001	.298	.878	yes
	(.17)	(.144)	(.179)		(.161)		
School variables							
School size (no. students)	581.171	597.592	573.115	.461	623.665	.950	yes
	(325.934)	(293.526)	(340.95)		(270.55)		
Prop. of dropout	.103	.126	.092	.006	.119	.682	yes
	(.114)	(.116)	(.118)		(.118)		
Prob. of dropout in	.23	.3	.196	.028	.327	.693	yes
high quartile(=1)	(.422)	(.46)	(.398)		(.47)		
ESCS in high quartile(=1)[d]	.248	.146	.298	.000	.16	.931	yes
	(.432)	(.355)	(.458)		(.367)		
Student-Teacher Ratio	9.441	9.091	9.612	.347	9.367	.648	yes
	(7.048)	(2.159)	(8.471)		(2.2)		
Parental pressure[e]	.357	.369	.351	.724	.397	.920	yes
	(.48)	(.484)	(.478)		(.49)		
School climate-teacher[f]	.554	.669	.498	.001	.712	.606	yes
	(.498)	(.472)	(.501)		(.454)		
Observations	395	130	265		265		

Standard deviations in parentheses.

[a] Initial test score corresponds to the average score in the first five questions of the first cluster of the test.

[b] The index of education possession indicates whether the home possesses a desk and a quiet place to study, a computer and/or educational software and books to help with school work, and a dictionary. It ranges between -3.93 and 1.12.

[c] The mother is defined as highly educated if she has achieved at least tertiary education.

[d] Index of economic, social, and cultural status.

[e] The dummy is equal to 1 if the principal claims that parents exert pressure into teachers and principal to improve the school quality.

[f] It is a dummy equal to 1 if the school is below the median value of the index of teacher-related factors affecting school climate. Positive values indicate that the teacher-related behaviors hinder learning to a lesser extent. The index ranges between -3.2778 + 2.8533.

Here we consider the school as the unit of analysis. Before estimating the impact of the PAE on outcomes, we take average of all variables, that is, we *collapse* the data at the school level. We then proceed as in the student analysis above: we estimate the probability of participating in the PAE (the propensity score), use the estimated propensity score to construct the re-weighted sample of control schools, and we use the previous results to compute the inverse probability weighting estimator (with and without

covariates). As above, we also provide results for the simple OLS. Notice that, for the impact of the PAE on rate of decline, we used weighted averages taking into account the school sample size. School characteristics are comparable between treated and re-weighted sample of control schools, as reported in Table I.1.

The outcomes considered are the mean school rate decline and the percentage of students at school with rate of decline in the first quartile of the rate decline distribution. Results can be found in Table I.2.

Table I.2: The impact of PAE on the rate of decline at the school level

	OLS		IPWE	
	(1)	(2)	(3)	(4)
	Level			
PAE	0.020	0.039*	0.043*	0.039*
	(0.021)	(0.022)	(0.023)	(0.022)
	P25 of the entire sample			
PAE	-0.010	-0.015	-0.017*	-0.017*
	(0.009)	(0.009)	(0.009)	(0.009)
Controls	No	Yes	No	Yes
Observations	395	395	395	395

Robust standard errors clustered at the school level in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For the estimation of the propensity score we use the following variables: gender, immigration status, grade repetition, attendance of the kindergarden, initial test score, index of education possession, mother's education, school size, probability of dropout, a dummy equal to 1 if ESCS is in the high quartile, student-teacher ratio, parental pressure on teachers, and school climate-teacher.

As it can be observed, they are very similar to those in Table 5 in the main text when considering the student as the unit of analysis. The effect of the program on rate of decline is between 0.039 and 0.043 of one standard deviation (compared to the 0.041-0.047 interval for the increase at the student level). We find that the percentage of students in the first quartile of the rate of decline distribution declines by 1.7 p.p. (compared to the 2 p.p. reduction at the student level). To conclude, results at the school level are in line to those at the student level.

J: Difference in Difference

We use the information from both the PISA 2009 and 2012 database and estimate the following model:

$$Y_{it} = \beta_0 + \beta_1 PAE_{it} + \beta_2 Post_{it} + \beta_3 PAE_{it} * Post_{it} + \delta_1 X_{it} + u_{it} \quad (5)$$

The outcome variables are the student i rate of decline, and whether she belongs to the first quartile in the rate of decline distribution using the complete questionnaire in the year t they are observed. PAE_{it} is the treatment variable which is a dummy equal to one if the student is attending a school with PAE in 2011/12 regardless of whether it had it before or not (but always after 2009). $Post_{it}$ is a dummy variable equal to one for students' outcome observed in PISA 2012 and equal to 0 for students' outcome observed in PISA 2009. Finally $PAE_{it} * Post_{it}$ is the interaction term between the dummy for treatment status of the school and year. Thus, the coefficient of interest (β_3) is the difference-in-difference estimator that captures the difference in outcomes between the treatment and control schools, before and after the introduction of the program. X_{it} denotes the vector of pre-treatment characteristics for students. Robust standard errors are clustered at the school level, as in the main analysis.

There are 137 schools that participated in both PISA 2009 and PISA 2012 (excluding private schools and schools which joined other remedial programs). Our sample here consists of 6,558 students of those schools: 1,335 students in 30 treated schools and 5,223 students in 107 control schools. The school sample reduces by 65% (from 395 schools in our main analysis to 137 now). Table J.1 shows the characteristics of students in the main analysis (column (1)) and the characteristics of students in this alternative analysis in the common year 2012 (column (2)). As it can be observed, it differs in some characteristics from the sample used in the main analysis. For instance, those schools in 2012 that also participated in 2009 have fewer repeaters, dropouts and migrants, a larger proportion of educated parents, higher ESCS index and lower teacher-student ratio.

Table J.1: Summary statistics

Variable	(1) Main Analysis	(2) DiD	(3) P-value Diff. (2)-(1)
Individual variables			
Initial test score[a]	.621	.639	.001
	(.272)	(.269)	
Girl(=1)	.506	.502	.730
	(.5)	(.5)	
Migrant(=1)	.107	.101	.268
	(.309)	(.301)	
Repeated once(=1)	.237	.213	.003
	(.425)	(.409)	
Repeated more than once(=1)	.087	.066	.000
	(.282)	(.248)	
Attended kindergarden(=1)	.839	.825	.061
	(.367)	(.38)	
Socioeconomic variables			
Index of education possession[b]	.063	.079	.374
	(.885)	(.888)	
Mother highly educated(=1)[c]	.345	.393	.000
	(.475)	(.489)	
Father highly educated(=1)[d]	.33	.384	.000
	(.47)	(.487)	
School variables			
School size (no. students)	606.893	614.722	.251
	(318.336)	(349.858)	
Prop. of dropout	.095	.076	.000
	(.109)	(.103)	
Prob. of dropout in	.236	.412	.000
high quartile(=1)	(.425)	(.492)	
ESCS[e]	-.274	-.171	.000
	(.977)	(.945)	
ESCS in high quartile(=1)	.272	.336	.000
	(.445)	(.472)	
Student-Teacher Ratio	9.621	8.574	.000
	(7.213)	(3.815)	
Parental pressure on teachers(=1)[f]	.356	.3	.000
	(.479)	(.458)	
School climate-teacher(=1)[g]	.564	.535	.003
	(.496)	(.499)	
Rural(=1)[h]	.42	.427	.453
	(.494)	(.495)	
Observations	11,105	3,310	

Standard deviations in parentheses.

[a] Initial test score corresponds to the average score in the first five questions of the first cluster of the test.

[b] The index of education possession indicates whether the home possesses a desk and a quiet place to study, a computer and/or educational software and books to help with school work, and a dictionary. It ranges between -3.93 and 1.12.

[c] The mother is defined as highly educated if she has achieved at least tertiary education.

[d] The father is defined as highly educated if he has achieved at least tertiary education.

[e] Index of economic, social, and cultural status.

[f] The dummy is equal to 1 if the principal claims that parents exert pressure into teachers and principal to improve the school quality.

[g] It is a dummy equal to 1 if the school is below the median value of the index of teacher-related factors affecting school climate. Positive values indicate that the teacher-related behaviors hinder learning to a lesser extent. The index ranges between -3.2778 + 2.8533.

[h] It is a dummy equal to 1 if the school is located in a village or a small town.