Test performance and Remedial Education: Good news for girls[*]

Marianna Battaglia
University of Alicante[†]

Marisa Hidalgo-Hidalgo
University Pablo de Olavide[‡]

June 15, 2018

Abstract

Growing evidence shows that skills other than cognitive are crucial to understand labor market and other outcomes in life. However, little is known about the role of education in improving these other abilities for disadvantaged students in developed countries. In this paper we evaluate the effects of a remedial education program for under-performing students from poor backgrounds implemented in Spain between 2005 and 2012. We address the following questions: (i) Does the program improve skills in test taking regardless of cognitive skills? (ii) Can we expect heterogeneous effects depending on the students' gender? We use external evaluations of the schools (PISA 2012) and exploit the variation in the question ordering of the test to compute students' ability to sustain performance throughout it. Our findings suggest that the program had a positive effect on girls' test performance. We find no impact for boys.

Keywords: remedial education, test performance, program evaluation, PISA

JEL classification codes: H52, I23, I28, J24

[*] We are grateful to Kevin Denny, Paul Devereux, Sue Dynarski and Chris Jepsen for helpful comments and suggestions. We benefited from valuable comments of seminar participants at 2018 UCD Workshop on the Economics of Education and Alicante Workshop on Quantitative Economics. Financial support from Fundación Ramón Areces is gratefully acknowledged. All opinions expressed are of the authors, all errors are our own.

[†] 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, Sevilla, Spain. Tel: +34 954 977 982. Contact email: mhidalgo@upo.es

1 Introduction

In skill acquisition both cognitive and non-cognitive abilities are relevant and equally important in explaining long-term outcomes such as high education investment and job market perspectives. As suggested by a growing body of the literature, skills other than cognitive are as crucial as cognitive skills in determining students' school achievements and in turn their educational choices.[1] Moreover, and even more relevant to our study, as suggested by Carneiro and Heckman (2003), both types of skills differ in their malleability over the life cycle, with non-cognitive skills being more malleable than cognitive 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. However, the effect of remedial education programs on non-cognitive abilities has so far been rarely investigated.[2] In this paper, we attempt to address the following questions: (i) Does a remedial education program improve skills in test taking regardless of cognitive skills among students with poor achievements? (ii) Can we expect heterogeneous effects depending on the students' gender? We do so by evaluating the effects of a multiyear program implemented in Spain between 2005 and 2012 that offered remedial education for under-performing students from poor socioeconomic backgrounds. This remedial program is the Program for School Guidance (PAE).[3] Similar to recent literature, we use the term non-cognitive skills to describe the personal attributes not thought to be measured by IQ test or the like. Specifically, we consider testing and survey behavior, for instance decline in test performance, as measures of non-cognitive skills (see Balart et al. (2018), among others, and the literature reviewed below). Data are obtained from external evaluations of the schools, the PISA 2012 tests, and we exploit the variation in the question ordering of the test to compute students' sustained performance during it.

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 Europe as there is less of a tradition compared to U.S. where remedial education is quite widespread (see references

[1] See, among others, Heckman and Rubinstein (2001), Heckman et al. (2006), Cunha and Heckman (2008), Carneiro et al. (2007) or Lindqvist and Westman (2011).

[2] An important exception is Heckman (2000) who provide a complete review on evidence on several interventions in adolescents during the nineties in the US.

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

mentioned above). However, 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. Students' individual and socioeconomic characteristics affect both their probability of being selected for the program and its success, when the selection mechanism is not completely observable. Only a few works address the identification problem and obtain usually positive evidence regarding the effectiveness of these programs in the short run. We comment below how this paper departs from previous works and contributes to the literature.

Our estimation strategy compares skills in test taking 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. The counterfactual outcomes are inferred using a control group composed of students in schools that did not join the PAE but participated in PISA 2012. 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 those in the treated group.[4] Since we cannot observe whether a particular student is actually treated, to obtain a more precise estimation of the true effect of the program, we also decompose our evaluation sample and focus on students who are more likely to participate: those whose parents have a low education level and those enrolled in schools with a high proportion of immigrants and repeaters. In addition, we replicate our main analysis using the school as unit of interest.

The main findings of the paper suggest that the PAE has a substantial positive effect on students' sustained test performance: it reduced the probability of falling behind into the bottom part of the rate of decline in test performance distribution by about 2 percentage points. The estimated increase on mean rate of decline in test performance is between 0.041 and 0.049 of one standard deviation. The corresponding figures (reduction in the probability of falling behind the bottom part and increase in mean rate decline) for girls are 4.6 percentage points and 0.1 of one standard deviation. We found no impact of the program on boys. The estimated impact of the program for the sub-sample of students with higher chances of being treated (at schools with a high proportion of migrants and repeaters) is quite similar in size to the impact for the whole sample of students, thus suggesting that we come close to estimate the true impact of the program when using the whole sample. Finally, our results hold when we consider the

[4] See also García-Pérez and Hidalgo-Hidalgo (2017) for a the same empirical strategy or Hospido et al. (2015) who employ a similar approach to examine the impact of financial education program on student' scores.

school, instead of the student, as the unit of analysis. We are also interested in investigating how other specific non-cognitive skills are correlated with test taking, in line with the analysis of Balart and Oosterveen (2018). We consider students' self-assessed measures such as absenteeism and truancy, discipline measured by the way students behave in class, self-confidence, sense of belong to the school, and perception of learning at schools (Supplementary Material Section F). 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 PAE and presents the data and descriptive statistics used in the paper. Section 5 describes the methodology. Section 6 reports the results. Section 7 discusses the validity of results and Section 8 concludes.

2 Brief literature review

Our paper contributes to three strands of the literature: the evaluation of remedial 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) evaluate the short-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. Non-cognitive skills were the objective of a remedial education program studied by Holmluld and Silva (2014). Such program targeted English secondary school pupils at risk of school exclusion and has been found to have little effect in helping treated youths to improve their age-16 test outcomes.[5] The most closely related papers to our are Battaglia and Lebedinski (2015) and García-Pérez and Hidalgo-Hidalgo (2017). The former analyzes the impact of the Roma Teaching Assistant program in Serbia, the main intervention targeting Roma inclusion in education in South Eastern Europe, on cognitive and non-cognitive skills. They find an overall positive effect of the remedial education program: children exposed to it are less absent from school. Moreover, first graders report lower dropout rates and better marks. García-Pérez and Hidalgo-Hidalgo (2017) analyze

[5] Additionally, 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) analyze 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.

the same remedial program as in this paper but focus on cognitive skills, measured by PISA test scores. They find that PAE had a substantial positive effect on children's academic achievement and that a larger exposure to the program improves students' scores. This paper departs from the previous works by evaluating the impact of a remedial education program beyond cognitive skills and focuses on abilities proved to be more likely affected by policy interventions at later stages of one persons' life, as remedial education programs are. In the paper we go through each and every single question in the PISA test which allows us to elicit students' testing behaviors.

This paper also provides new insights on recent works on skills other than cognitive. Borghans and Schils (2012) use 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, and show that it is a good predictor of final levels of educational attainment, without being related to cognitive performance. Using 2009 PISA, Zamarro et al. (2016) expand the methods used by Borghans and Schils (2012) 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 program 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 (2018) 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.[6] Our findings confirm these conclusions and move forward

[6] For instance, Jacob (2002) shows that girls have less behavioral problems and Cornwell et al. (2013) found that girls show more developed attitudes towards learning, etc.

them by analyzing whether girls are not only better in non-cognitive skills than boys but also more apt to improve them when receiving remedial education.

3 Test performance and the PISA test

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; Holmluld 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). We build on previous research (see e.g. Borghans and Schils (2012); Balart and Oosterveen (2018); Zamarro et al. (2016) mentioned above) which uses students' response patterns to surveys and tests to get a non-self assessed measure for their personality traits. The idea is that students, through their effort on tests and surveys, might provide some information about their conscientiousness, self-control or persistence. Following recent literature, we exploit the variation in the question ordering of the PISA 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 cognitive abilities.

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 (see 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.

Here Table 1: Rotation design of the PISA booklets

As shown in Table 1 above, 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 students are allowed a short break after one hour. Clusters labeled *Math 1, Math 2, Math 3, Math 4, Math 5, Math 6A and 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.[7] 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. For instance, cluster Maths 5 is included in booklets 1, 5, 9 and 11 as respectively the first, forth, third and second cluster. This random assignment ensures that the variation in scores' decay is unrelated to characteristics of students. 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:

$$y_q = \alpha_0 + \alpha_1 p_q + \alpha_2 d_q + u_q \tag{1}$$

where y_q is a dummy for whether student i answered question q correctly, p_q 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 the difficulty of the question q (from simple choice to multiple choice or open question).[8] As our dependent variable is a dummy, we estimate a probit model. Our coefficient of interest is α_1 which shows the individual pattern of the test performance drop. A significant and negative (positive) coefficient would reveal a decline (improvement) in performance from the first to the last question of the test.[9]

As an alternative measure of student's non-cognitive skills, we also use item reached during the test, corresponding to the average last question answered by the student in each cluster. The summary statistics and results are reported in Section D in the Supplementary Material.[10]

In addition to Equation (1) we estimate three comparable models for rate of decline in per-

[7] Balart and Oosterveen (2018) 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.

[8] As an alternative definition of correct answer, we recode a question as correct if the answer is correct or partially correct. We also provide two different measures of difficulty: (i) a dummy variable equal to 1 if it is a simple question and 0 otherwise and (ii) the percentage of students who correctly answer the question. See Section 6 for comments on robustness of our main results to these alternative definitions.

[9] Balart and Oosterveen (2018) also check for the non-linearity effect of the position of the question finding similar qualitative results than under the linear assumption.

[10] The results are consistent to choosing the minimum or the maximum last question answered. They are not reported but are available upon request.

formance by considering the specific clusters of maths, reading, and science questions instead of the complete questionnaire in the PISA test. Table 3 below shows the (standardized) estimated rate decline for the complete PISA test, for maths, reading and science. It reports the values by gender and overall. It also reports values for treated, controls and the weighted control group (see below). A negative (positive) rate of decline measures the % reduction (increase) in the probability of correctly answering a question as the position of that question increase 1% from the first to the last question.

Figure 1 depicts the decline in performance during the test considering the complete questionnaire and the maths, reading and science clusters.

Here Figure 1: Decline in performance

Figure 2 reports the same information separately for boys and girls.

Here Figure 2: Decline in performance: the gender gap

Several comments can be made from this figure. First, the average estimated rate of decline is negative, in particular it is equal to -.097. That is, there is a decline in performance during the test which confirm previous findings by Borghans and Schils (2012); Balart and Oosterveen (2018); Zamarro et al. (2016). Second, the average estimated rate of decline is lower among girls, which is also in line with recent evidence by Balart and Oosterveen (2018) who find that girls have a higher ability to sustain performance. As it can be observed in the complete questionnaire, there is an initial gap in test scores favoring boys, however, during the test this advantage vanishes and girls finish the questionnaire outperforming boys. In the maths and science clusters boys outperform girls since the beginning of the test whereas girls score better than boys in the reading clusters. Finally it can also be observed that in the maths and science clusters the initial gap favoring boys reduces with the progress of the test. In the reading clusters the initial gap favoring girls increases during the test.[11]

As the main goal of the PAE was to improve poor educational outcomes among students from disadvantaged backgrounds, we concentrate our analysis on the performance of that specific group of students. We define the group of *low achievers* by using the score in the first quartile of each rate of decline distribution (for the complete PISA test and the maths, reading and science

[11] All gender differences are statistically significant at 0.01 level.

clusters). Additionally, we also consider as an outcome variable the student's decline in test performance. Thus, in the rest of the paper we focus on the following two outcome variables: (i) the probability of falling behind the general progress of the group or being a low achiever; (ii) each student's rate of decline.

4 The remedial program

As mentioned above, Carneiro and Heckman (2003) and Heckman (2000) provide evidence which suggest that non-cognitive abilities are more likely than cognitive ones to be affected by policy interventions at later stages of one persons life and can therefore be relevant when teenagers are involved in a remedial education program (as it is in our case).

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. This was pursued by stimulating reading habits, providing students with study organization techniques, and improving their social abilities. It consisted of providing support (at least 4 hours per week) during after-school hours to those students with special needs and learning difficulties. This support was provided in small groups of 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 secondary school. They were chosen based on their poor academic results, general motivation and prospects, although there was no single quantifiable and explicit selection rule. 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, assisted students with work organization techniques, etc.

The PAE was implemented during the period 2005-2012. It provided support to public schools with a significant number of students from disadvantaged backgrounds. The PAE was progressively introduced throughout the period 2005-2012. The percentage of schools participating in the PAE was very low during the first three academic years (below 1% in most regions), while it started to be gradually implemented in most regions during the 2008-2012 period.[12] We focus here on 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

[12] See Figure A in Section A in the Supplementary Material.

secondary schools where they took the PISA exams. The reason is that, during the 2005-2008 period even the schools participated in the program, students in our sample did not benefit from PAE since they were attending primary school.[13]

The PAE 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 and committed themselves to improve their students' outcomes by providing after-school instruction to those students with special needs. They received funding from the regional authorities and had to manage program implementation. 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.

Even though PAE was implemented in both primary and secondary schools, we focus our analysis on secondary schools. The reason is that PISA 2012 exam is taken by 15-year-old students, with 10th grade being the reference grade for them. Moreover, as the program was implemented only in public schools, we exclude from the PISA database both private and private but publicly financed schools.[14] Following García-Pérez and Hidalgo-Hidalgo (2017), we do not consider in the analysis schools that joined other remedial programs.[15] Our sample consists of 11,105 individuals from 395 schools, corresponding to 44% of the Spanish schools in PISA 2012 database (with 902 schools).[16]

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 2005/06 and 2010/11).[17] We consider as *controls* students in schools where the PAE was not implemented

[13] 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). She starts school at 6 years old. Most schools provide either primary or secondary and pre-college education. Only a very small sample of schools (most of them private) provide the three levels. See Spanish Ministry of Education (2016).

[14] We excluded 352 schools because they are private or private publicly financed schools.

[15] From the initial 550 public schools in PISA 2012 database we exclude 133 schools because they participated in other remedial programs.

[16] We drop from the analysis 622 students in 22 schools where the PAE was implemented during any academic year between 2005/06 and 2010/11 but not thereafter, i.e., during 2011/12.

[17] Alternatively we could analyze 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 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.

at all (that is, in no academic year between 2005/06 and 2011/12). As a result, there are 129 treated schools (with 3,660 students) and 266 control schools (with 7,445 students) in our sample.[18]

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), school-level variables and achievement test scores in three disciplines: maths, reading and science.

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 the treated students (column (2)), control students (column (3)) and the differences between them (column (4)).

Here Table 2: Summary Statistics

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 treated students, suggesting that treated schools have a higher proportion of students from disadvantaged backgrounds. Initial test score, measured as the average test score in the first five questions of the first cluster of the test, is nonetheless not statistically significantly different between the two groups. Finally, treated students came from larger sized schools that exhibited a larger proportion of dropouts and lower ESCS. Conversely, students in the control sample are from schools with a higher student-teacher ratio, where principal enhance school's reputation and parents exert less pressure on teachers. They are also at schools with less migrants. In the analysis below, we comment on weighted control students in column (5)

[18] There are at most 35 students per school participating in PISA. These students are selected based on a two-stage sample design developed by the PISA program organizers. This selection ensured representation of the full target population of 15-year-old students in the participating countries. Only in a few cases, and with proper justification, PISA national project managers can exclude certain schools (e.g., in a remote geographical region) or students (e.g., special needs students). Nevertheless, the guidelines explicitly state that students must not to be excluded solely because of poor academic performance or normal discipline problems. See the PISA 2012 Technical Report for further details on PISA 2012 and García-Pérez and Hidalgo-Hidalgo (2017) for details on how the PAE was introduced in the schools.

of Table 2 and on the difference between the treated and the weighted control group.

Table 3 below presents the rate of decline by gender for treated and control students.

Here Table 3: Students' outcomes: rate decline

Observe that, the percentage of boys in the *poorest skilled* group (first quartile of the rate decline distribution) is larger than the percentage of girls.[19]

5 The empirical strategy

We study the effects of the PAE on the student's rate of decline in test performance and on her probability of falling behind the general progress of the group (having a rate decline in the first quartile of the rate decline distribution). To the extent that we cannot observe whether a particular student actually received the treatment, by selecting the student as the unit of observation, we are aware that we can only consider her *potentially* treated. Nevertheless, we address this point below and attempt to provide a cleaner estimate of the true effect of the PAE by decomposing our evaluation sample. In addition, we study the impact of the program while considering the school to be the treatment unit.

In the evaluation literature, data often come from non-randomized studies. The main assumption is that individuals' participation in the policy intervention can be considered a random event or, at least, independent of treated and control individuals' characteristics (see Myoung-JaeLee (2005)). However, selection into the treatment is not independent of treated and control individuals' characteristics. Propensity score matching is a method to reduce the bias in the estimation of treatment effects when using such datasets. The propensity score is defined by Rosenbaum and Rubin (1984) as the probability of being treated considering those variables included in the set of regressors.The method proposes to summarize the pre-treatment characteristics of each subject into a single-index variable (the propensity score) that makes the matching feasible. This index is built based on the estimation of the probability of being treated, $p(X_i)$, where X_i denote the vector of pre-treatment characteristics. If D_i denote a binary variable that indicates exposure to the treatment:

[19] This result holds for the science specific clusters (interestingly, those in which boys tend to perform better than girls) but not in the reading cluster (in which girls tend to outperform boys). Results available upon request.

$$D_i = \begin{cases} 1 & \text{if treated} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

the propensity score is defined as the probability of PAE participation conditional on some pre-treatment characteristics, X_i:

$$p(X_i) \equiv Pr(D_i = 1|X_i) = E(D|X_i) \tag{3}$$

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. We denote by Y_i the outcome (rate of decline or probability of falling into the first quartile of the rate decline distribution), where $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$. Therefore, the average effect we are interested in estimating when evaluating the PAE is

$$\tau = E(Y_i^1|D = 1, X_i) - E(Y_i^0|D = 1, X_i) \tag{4}$$

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 that the characteristics of the control and treatment group be as similar as possible. In our sample, as previously mentioned, treated and control students differ in their demographic characteristics, in socioeconomic background and attend different schools (see Table 2). To solve 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 scores of the treated students. Formally, under the standard assumptions of conditional independence or unconfoundedness:

$$(Y_i^1, Y_i^0) \perp D_i \mid X_i \tag{5}$$

that is, within each cell defined by X_i, treatment is random, or similarly, the selection into treatment depends only on the observables X_i, and common support:

$$p(X_i) \in (0, 1), \tag{6}$$

we have that:

$$E(Y_i^0|D=1,X_i) \equiv E(\omega(x_i)Y_i|D=0,X_i) \tag{7}$$

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

This expression indicates that we can identify the mean impact on treated individuals were they to have not received the treatment, $E(Y_i^0|D=1,X)$, by re-weighting the sample of controls. Observe that the weights, $\omega(x_i)$, increase the relevance in the control sample of those individuals who are very similar to treated students, 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, $p(X_i)$. We therefore compute the *inverse probability weighting estimator (IPWE)*. This estimator is achieved by regressing the outcome variable (either the rate of decline or the probability of falling behind the lowest quartile) on the treatment, where each observation is weighted by $\omega(x_i)$.[20] Since, through the consideration of the propensity score in the weighing 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. Since we observe that boys and girls differ in non-cognitive skills, we also analyze whether they equally benefitted from the program by adding an interaction term for the treatment and student's gender.

Finally, we comment on the validity of the two assumptions we make: unconfoundedness and common support. If the first assumption is not satisfied, this means that program participation could be due, among other reasons, to special interest by parents, teachers or school principals. If these variables are positively correlated with the distribution of potential outcomes (i.e., more interested parents or teachers are also more likely to yield better student non-cognitive skills), then our estimates of the impact of the PAE would be biased; in particular, they would be overestimating the true impact of the program. However, these unobserved school characteristics might also be negatively correlated with students' outcomes, for example, the existence of a difficult student body at the school. In that case, then our previous results would be underestimating the true impact of the program. This assumption is therefore crucial. We attempt to address it by including a set of variables that might capture these parent, teacher and school characteristics (particularly, the school ESCS and whether teachers affect school climate).[21]

[20] See García-Pérez and Hidalgo-Hidalgo (2017) and Hospido et al. (2015) for a similar approach and Hirano et al. (2003) or Busso et al. (2014) for methodological details.

[21] García-Pérez and Hidalgo-Hidalgo (2017), using the PISA 2009 dataset to characterize possible selection bias, show that no selection bias exists. They find that, if any, possible differences can be explained by differences

The second assumption, the common support, can be tested by comparing the propensity score densities of the treated and control groups. We check this assumption graphically in Figure 3. As it can be observed, the common support assumption seems to hold in our sample. 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 treated students (solid line) and that of the control (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).

Here Figure 3: Propensity score support

5.1 Participation in the remedial program

We estimate the predicted probability of participation in the remedial education program (PAE) 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. 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. Excluding such variable from the analysis does not change the results.[22] 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. Finally, we also add a set of school characteristics, including the student-teacher ratio, its mean socioeconomic index, its size, the proportion of dropouts, and an indicator of whether teachers favor good school climate. We then augment the basic logit model by including interactions that were statistically different from zero according to a two-sided t-test. This set of variables might affect the probability of participating in the program according to differences in mean covariates in Table 2 as commented above. The final specification is shown in Table 4. The first column presents the estimates of the propensity score for the treatment. Its weights are used to estimate the impact of PAE on the general rate of decline of the complete questionnaire. Columns (2)

in individual, parental and school characteristics. Accounting for these differences completely attenuates the selection bias. Therefore, this suggests that it is feasible to obtain estimates of the impact of PAE participation on non-cognitive skills with no selection bias by re-weighting the sample according to student, family and school characteristics, as we do.

[22] Results are not reported but they are available upon request.

to (4) present the estimates of the propensity score for the treatment whose weights are used to estimate the impact of the program on the rate of decline of maths, reading and science questions, respectively. As it can be observed, the specifications of the four propensity scores are the same.[23] This allows us to obtain comparable results across the different treatments.

Here Table 4: Propensity score estimation- Probability of being treated

The estimates in the first column confirm the results of Table 2. The mean initial test score at the school level does not affect the probability that the school offer the program. The proportion of boys (girls) in a school does not seem to affect the likelihood that a school joins the program. 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. Regarding school variables, those schools with poorer socioeconomic index, larger size, an a larger index of school climate have a higher chance of being treated. Finally, observe that the results of the propensity score when we consider the whole questionnaire are very similar to the ones obtained when we desegregate in the three specific questionnaires: maths, reading and science.

To conclude, column (5) of Table 2 presents the means of the control sample once the latter is re-weighted by $\omega(x_i) = \frac{1-\pi}{\pi} \times \frac{p(X_i)}{1-p(X_i)}$.[24] Column (6) column in Table 2 reports the differences in characteristics between treated and re-weighted controls. 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).[25] 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

[23] The only differences are that *migrant* for the reading questionnaire and *repeated more than once* for the science questionnaire do not satisfy the balancing property.

[24] Therefore 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. In particular there are 65 observations for which the estimated propensity score is missing. Observe that this is exactly the difference between the controls observations in column (3) in Table 2, 7,445 and the weighted controls observations in column (5), 7,375 in the same table.

[25] Exceptions are the proportion of migrants, parental pressure on teachers and principal enhancement of school reputation. The latter is lower in the treatment group while the others are lower in the control group.

treated.[26]

6 Main results

In this section, we comment on the impact of the program on students' test performance.

6.1 On the general effect of the program

The estimated general effect of the program on the rate of decline in the complete questionnaire is reported in Table 5. It presents the estimated impact of the treatment on mean rate of decline for the complete questionnaire, and the maths, reading and science specific questionnaires, and on the probability of belonging to the first quartile in the rate of decline distribution. Recall that we control for students' initial test score.[27] The rate of decline is standardized with the average and standard deviation of the sample of students in the complete, maths, reading and science questionnaires, respectively.

Here Table 5: The impact of PAE on Rate decline

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 in almost all cases. 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. This result can also be inferred from the first row in Table 3. As it can be observed there, the rate of decline among the treated is equal to .021, while that of the re-weighted control group is equal to -.028. The .049 difference is the observed impact of the program. The standard error accounts for arbitrary correlation at the school level and is equal to 0.024; thus, the estimate is statistically significant at the 10% 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 again 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. In addition, we go further and compare each treated

[26] See Lavy and Schlosser (2005) or Hospido et al. (2015) for a similar test.

[27] Excluding such variable from the analysis does not change the results (available upon request).

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 row 5 to row 8). As it can be observed, the results are quite similar to those obtained by using the inverse probability weighting estimator. In particular, the larger the number of nearest neighbors used, the more similar the results are to the IPWE ones. To summarize, we find that the program improved mean rate of decline by between 0.041 and 0.049 of one standard deviation.[28]

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 and columns (5) to (8) results using the nearest neighbor propensity score maching. The results are exactly the same when re-weighting estimates are used without and with covariates and are consistent with previous findings. As before, the result in column (3) can also be inferred from Table 3. The program reduced the probability of belonging to the bottom quartile in the complete questionnaire distribution by 2 p.p.

The rest of panels show the impact of the program using the maths, reading and science specific questionnaires. We do not observe statistically significant results, although they mainly go in the same direction as for the complete questionnaire.[29]

6.2 On the impact of the program by gender

Finally we analyze the impact of the program by student's gender. Results can be found in Table 6 below. Columns (1) to (4) show the estimated impact of the program on boys and columns (5) to (8) for girls.

Here Table 6: The impact of PAE on Rate decline by gender

[28] We also used alternative definitions of correct answer and difficulty of the question. Main results in the paper are robust to these other definitions. See Section B in the Supplementary Material.

[29] In addition we checked 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 between boys and girls. Results show that it is not the case here (see Section C in the Supplementary Material).

The rate of decline increases about 0.10 of one standard deviation, only for girls. We find no 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. Note that the positive impact of the program on the rate of decline in the complete questionnaire is less precisely estimated when we analyze the maths, reading and science questionnaire separately. A plausible explanation could be the reduced number of observations in those analyses. The program participation also reduced the probability of belonging to the bottom quartile only among girls. As before, the result in column (7) can also be inferred from Table 3. The proportion of treated girls in the first quartile in the rate of decline distribution is equal to .194, while that of the re-weighted control group is equal to .24. The -0.046 difference is the observed impact of the program. That is, the program participation reduced the probability of belonging to the bottom quartile by 4.6 p.p. among girls. Observe also from table 8 that the probability of falling behind into the bottom part of the distribution is reduced by 4 p.p. for those girls in the sub-sample of schools with immigrants or repeaters above the median and by 4.2 p.p. for those girls in the sub-sample of students in non-educated families. When focusing on maths, reading and science specific questionnaires we find that the program reduced the probability of belonging to the bottom quartile in the maths questionnaire by between 3.2 and 2.5 p.p. again only among girls. No robust effects are observed for the reading and science questionnaires. Therefore, we can conclude that the observed reduced probability of belonging to the bottom quartile in the rate of decline distribution for the complete questionnaire might be mostly driven by the impact on the maths specific questionnaires.

6.3 Discussion

In this section we investigate the potential mechanisms explaining the impact of the program only on girls. First, girls could be over-represented in those percentiles in the rate decline 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. We calculate the values of two Cumulative Distribution Function (CDF) of the rate decline for certain percentiles: the CDF of rate decline among treated students and 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 4 below shows the results.

Here Figure 4: The impact of PAE: Gender

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 tails of the distribution, precisely 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 7.

Here Table 7: Summary statistics by gender

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 an higher index of education possession. If we were expecting a differential participation to the program by gender, boys could participate more than girls to it.

Third, girls might participate for longer (less attrition) and they can better respond to the PAE. The remedial education program is more effective in improving skills other than cognitive for girls.

7 Further discussion on impact of the program

As previously noted, the results for the full sample presented above might not precisely capture the true impact of the PAE. On the one hand, 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. On the other hand, by considering all of the students in the PAE school as treated, we might be capturing peer effects of treated on non-treated students. This assumption can lead to an overestimation of the impact of the PAE on treated students. We first decompose our evaluation sample. Second, we replicate the previous analysis considering the school as the unit of analysis.

7.1 Sub-sample analysis: disadvantaged students

To argue that the effect analyzed is close to the actual effect of the intervention on treated students, we focus our main analysis on two sub-samples of our treated students group. In particular, we split that group according to some pre-treatment characteristics, namely the proportion of migrants and repeaters at the 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 as they do not satisfy the balancing property. 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 treated students at schools where the proportion of migrants and repeaters 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 treated students with non-educated parents.[30] Table 8 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 of decline distribution. Rows (2) to (4) provide results for the sub-sample of students at schools with a proportion of immigrants and/or repeaters above the median. Rows (6) to (8) provide results for the sub-sample of students in non-educated families.

Here Table 8: The impact of PAE on Rate decline (Subgroups)

The estimated impact of the program on the rate of decline is an increase of 0.041 of one standard deviation, in the sub-sample of schools with migrants or repeaters above the median, which is very close to the impact on the full sample of students (between 0.044 and 0.044). Therefore, by considering the full sample of students, we came close to estimating the true

[30] In this analysis only treated students 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 quite similar to the ones found here. This is because control students at schools with a proportion of immigrants and repeaters above the median might not be that similar to treated students 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. Results are robust to different definitions of *non-educated* such as parents' level of education lower than lower secondary education or lower than primary education.

impact of the PAE on moving students out of low-skills status, which is the main objective of the program. The probability of belonging to the bottom quartile is reduced by 2.1 p.p overall, when considering schools with migrants or repeaters 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 or repeaters above the median and with the full sample, but standard errors are bigger and the number of observations reduces by 35%.

7.2 On the impact of the program at the school level

Next, we consider the school as the unit of analysis. Recall that to the extent that we cannot observe whether a particular student actually received the treatment or not, we might not capture the true effect of the PAE. Therefore the analysis at the school level is crucial. 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 a simple OLS model (with and without covariates) and the inverse probability weighting estimator (with and without covariates). Notice that, for the impact of the PAE on mean 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 A in Section E. 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 (P25).

Here Table 9: The impact of PAE on Rate decline at the School level

Results can be found in Table 9. As it can be observed, the results are very similar to those in Table 5 when considering the student as the unit of analysis. The effect of the program on mean rate of decline is 0.038 of one standard deviation (compared to the 0.041 increase at the student level). We find that the percentage of students in the first quartile of the rate of decline distribution declines by 1.79 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.

8 Concluding remarks

There is ample evidence of increasing inequality and poverty figures in developed countries.[31] This 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. In fact, one of the EU's education targets for 2020 is to reduce the rates of young people leaving early the education and training systems. National governments are currently being encouraged to undertake evidence-based education policies to reduce the adverse effects of the aforementioned facts. Surprisingly, it is difficult to find empirical evidence regarding the effectiveness of most of these interventions and in particular remedial education programs. In this paper, we estimate the effects of a remedial program implemented in Spain between 2005 and 2012 that offered additional instruction time for underperforming students from poor socioeconomic backgrounds: the Program for School Guidance (PAE). We concentrate on skills other than cognitive since they proved to be more likely affected by policy interventions at later stages of one persons' life, as remedial education programs are. Our main finding is that this program had a substantial positive effect on students' test performance. In particular, it helps 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 4.6 p.p. and reduces the decline in performance during the test by 0.10 of one standard deviation. We found no impact of the program among boys.

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 aiming at 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

[31] Recent evidence (OECD, 2013) suggests increases in inequality and poverty. This might be caused by the global crisis and might also reflect the fact that as a result of rapid technological change both low-skilled workers and low-achieving students are being left behind (see Freeman (2008) or Kanbur (2014)). Indeed, 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. Their disadvantage on the labour market is reflected in high unemployment rates, below average wages and possibly high concentration in the informal employment. They are poorer than the average population and more likely to fall into poverty and remain poor, with consequences in increased inequality.

comprehensive analysis of the strength of such programs.

References

Balart, P. and M. Oosterveen (2018). Wait and See: Gender Differences in Performance on Cognitive Tests. Working Paper.

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.

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. and T. Schils (2012). The leaning tower of Pisa: decomposing achievement test scores into cognitive and noncognitive components. Working Paper.

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)*, 885–897.

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

Carneiro, P., C. Crawford, and A.Goodman (2007). The Impact of Early Cognitive and Non-Cognitive Skills on Later Outcomes. *CEE Discussion Papers 0092, Centre for the Economics of Education, LSE.*

Carneiro, P. and J. Heckman (2003). Human Capital Policy. *IZA Discussion Papers 821, Institute for the Study of Labor (IZA).*

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.

Freeman, R. (2008). Globalization and Inequality. In B. N. W. Salverda and T. S. O. O. U. Press (Eds.), *The Oxford Handbook of Economic Inequality.* W. Salverda, B. Nolan and T. Smeeding. Oxford: Oxford University Press.

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.

Holmluld, 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 Espaa WP 1502 2015, IZA DP 8902 2015*.

Jacob, B. A. (2002). Where the boys aren't: Non-cognitive skills, retunrs to school and the gender gap in higher education. *Economics of Education Review 21(6)*, 589–598.

Kanbur, R. (2014). Globalization and Inequality. *Working Papers 180163, Cornell University, Department of Applied Economics and Management*.

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.

Myoung-JaeLee (2005). Micro-Econometrics for policy, Program, and Treatment Effects.

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.

Rosenbaum, P. and D. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*, 516–524.

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., C. Hitt, and I. Mendez (2016). When Students Dont Care: Reexamining International Differences in Achievement and Non-Cognitive Skills. *EDRE Working Paper No. 2016-18*.

Tables

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

Table 2: Summary Statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variable	All	Treated	Controls	P-value	Weighted	P-value	P-score
				Diff. (2)-(3)	Controls	Diff. (2)-(4)	
Individual variables							
Initial test score[a]	.621	.607	.628	.000	.604	.630	yes
	(.272)	(.274)	(.272)		(.277)		
Girl(=1)	.506	.5	.509	.383	.498	.864	yes
	(.5)	(.5)	(.5)		(.5)		
Migrant(=1)	.107	.15	.086	.000	.159	.214	yes
	(.31)	(.357)	(.281)		(.366)		
Repeated once(=1)	.237	.272	.22	.000	.273	.897	yes
	(.425)	(.445)	(.414)		(.445)		
Repeated more than once(=1)	.087	.106	.078	.000	.115	.154	yes
	(.282)	(.308)	(.268)		(.319)		
Attended kindergarden(=1)	.839	.83	.844	.064	.826	.581	yes
	(.367)	(.376)	(.363)		(.379)		
Socioeconomic variables							
Index of educ possession	.063	.041	.074	.065	.054	.483	yes
	(.885)	(.887)	(.884)		(.893)		
Mother highly educated(=1)	.345	.302	.366	.000	.301	.983	yes
	(.475)	(.459)	(.482)		(.459)		
Father highly educated(=1)	.33	.297	.347	.000	.298	.945	no
	(.47)	(.457)	(.476)		(.457)		
School variables							
Student-Teacher Ratio	9.621	9.286	9.785	.000	9.306	.684	yes
	(7.213)	(2.048)	(8.687)		(2.829)		
ESCS	-.274	-.374	-.223	.000	-.376	.901	no
	(.977)	(.972)	(.974)		(.951)		
ESCS in high quartile(=1)	.272	.149	.332	.000	.149	.956	yes
	(.445)	(.356)	(.471)		(.356)		
Prop. of dropout	.095	.115	.085	.000	.119	.135	yes
	(.109)	(.111)	(.107)		(.118)		
Prob. of dropout in	.236	.311	.199	.000	.304	.464	yes
high quartile(=1)	(.425)	(.463)	(.399)		(.46)		
School size	606.893	624.775	598.101	.000	632.083	.193	yes
	(318.336)	(277.686)	(336.194)		(277.997)		
Prop. of migrants (school)	.107	.15	.086	.000	.121	.000	no
	(.122)	(.144)	(.103)		(.139)		
Parental pressure	.356	.386	.341	.000	.342	.000	no
on teachers(=1)	(.479)	(.487)	(.474)		(.475)		
School climate-teacher	.564	.693	.501	.000	.7	.440	yes
	(.496)	(.461)	(.5)		(.458)		
Principal enhance	.229	.216	.236	.017	.276	.000	no
school's reputation(=1)	(.42)	(.411)	(.424)		(.447)		
Rural(=1)	.42	.405	.426	.035	.426	.037	no
	(.494)	(.491)	(.495)		(.495)		
Observations	11,105	3,660	7,445		7,375		

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.

Table 3: Students' outcomes: Rate decline

	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)
Complete questionnaire	-.053	-.084	-.038	-.073	.066	.125	.038	.018	.007	.021	.001	-.028
	(1.001)	(1.028)	(.991)	(1.029)	(.973)	(.954)	(.991)	(1.003)	(.989)	(.994)	(.986)	(1.017)
First quartile (P25)	.272	.286	.266	.281	.221	.194	.234	.241	.246	.24	.249	.261
Observations	5,429	1,823	3,606	3,606	5,581	1,827	3,754	3,754	11,010	3,650	7,360	7,360
Maths questionnaire	-.012	-.008	-.014	.005	.005	.031	-.007	-.007	-.003	.011	-.011	-.001
	(.984)	(.996)	(.978)	(1)	(.973)	(.993)	(.963)	(.976)	(.978)	(.994)	(.97)	(.988)
First quartile (P25)	.246	.253	.242	.235	.244	.222	.254	.262	.245	.237	.248	.25
Observations	5,146	1,726	3,420	3,420	5,341	1,749	3,592	3,592	10,487	3,475	7,012	7,012
Reading questionnaire	.038	.044	.033	-.007	-.012	-.03	0	.002	.012	.007	.016	-.002
	(.992)	(.989)	(.987)	(1.004)	(.973)	(.993)	(.958)	(.949)	(.983)	(.997)	(.973)	(.976)
First quartile (P25)	.237	.234	.239	.253	.248	.259	.239	.237	.242	.247	.239	.245
Observations	4,001	1,692	2,309	2,309	4,132	1,709	2,423	2,423	8,133	3,401	4,732	4,732
Science questionnaire	-.029	-.002	-.05	-.067	.038	.041	.036	.055	.005	.019	-.005	-.007
	(.991)	(.995)	(.987)	(1.006)	(.972)	(.958)	(.981)	(.99)	(.982)	(.977)	(.985)	(1)
First quartile (P25)	.266	.266	.267	.289	.236	.246	.23	.22	.251	.256	.247	.255
Observations	4,182	1,789	2,393	2,393	4,368	1,795	2,573	2,573	8,550	3,584	4,966	4,966

Standard deviations in parentheses.

Table 4: Propensity score estimation - Probability of being treated

	(1)	(2)	(3)	(4)
		Rate decline		
	Complete	Maths	Reading	Science
Individual variables				
initial test score[a]	0.085	-0.001	0.141	-0.046
	(0.095)	(0.103)	(0.138)	(0.102)
girl	-0.033	-0.040	-0.017	-0.119**
	(0.039)	(0.039)	(0.050)	(0.050)
migrant	0.458***	0.469***	-	0.488***
	(0.148)	(0.149)		(0.150)
repeated once	0.141**	0.129**	0.215***	0.004
	(0.061)	(0.061)	(0.070)	(0.063)
repeated more than once	0.179*	0.166	0.208*	-
	(0.103)	(0.105)	(0.123)	
attended kindergarden	-0.053	-0.048	-0.145	-0.052
	(0.100)	(0.100)	(0.108)	(0.102)
Socioeconomic variables				
mother highly educated	-0.027	-0.026	-0.040	-0.060
	(0.078)	(0.078)	(0.087)	(0.086)
index of education	0.005	0.009	0.006	-0.036
possession	(0.033)	(0.034)	(0.037)	(0.037)
School variables				
student teacher ratio	-0.034	-0.035	-0.036	-0.032
	(0.022)	(0.024)	(0.025)	(0.021)
ESCS[b]	-1.059***	-1.064***	-1.160***	-1.106***
	(0.332)	(0.333)	(0.337)	(0.334)
prob. dropouts[c]	0.236	0.372	0.375	0.366
	(0.331)	(0.270)	(0.270)	(0.268)
school size	0.004**	0.004**	0.004**	0.004***
	(0.002)	(0.002)	(0.002)	(0.002)
school climate-teacher	0.646**	0.651**	0.629**	0.656**
	(0.254)	(0.254)	(0.257)	(0.254)
Observations	10,975	10,958	7,335	7,541

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] ESCS is a dummy equal to 1 if the school belongs to the higher quartile of the distribution of ESCS at the school level.

[c] Probability of dropouts is equal to 1 if the school belongs to the higher quartile of the distribution of proportion of dropouts at the school level.

Table 5: The impact of PAE on Rate decline

	OLS		IPWE		NNPS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Complete questionnaire								
Level								
PAE	0.020	0.034	0.049**	0.041*	0.034	0.043*	0.044*	0.044*
	(0.021)	(0.023)	(0.024)	(0.023)	(0.027)	(0.024)	(0.023)	(0.023)
P25 of the entire sample								
PAE	-0.010	-0.015	-0.021**	-0.020**	-0.012	-0.014	-0.015	-0.017*
	(0.009)	(0.009)	(0.010)	(0.010)	(0.012)	(0.011)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	Yes	Yes	Yes	Yes
No. matches per obs.	-	-	-	-	2	4	6	8
Observations	11,089	10,964	11,010	10,964	10,964	10,964	10,964	10,964
Maths questionnaire								
Level								
PAE	0.020	0.011	0.013	0.013	-0.011	0.005	0.009	0.009
	(0.021)	(0.021)	(0.023)	(0.021)	(0.026)	(0.024)	(0.023)	(0.023)
P25 of the entire sample								
PAE	-0.011	-0.009	-0.012	-0.013	-0.008	-0.014	-0.014	-0.014
	(0.009)	(0.010)	(0.010)	(0.010)	(0.012)	(0.011)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	Yes	Yes	Yes	Yes
No. matches per obs.	-	-	-	-	2	4	6	8
Observations	10,570	10,437	10,487	10,437	10,437	10,437	10,437	10,437
Reading questionnaire								
Level								
PAE	-0.024	0.005	0.009	0.010	-0.006	-0.001	-0.002	0.000
	(0.029)	(0.025)	(0.030)	(0.025)	(0.032)	(0.029)	(0.028)	(0.028)
P25 of the entire sample								
PAE	0.014	0.003	0.002	-0.000	0.005	0.007	0.006	0.002
	(0.012)	(0.011)	(0.013)	(0.011)	(0.015)	(0.013)	(0.013)	(0.013)
Controls	No	Yes	No	Yes	Yes	Yes	Yes	Yes
No. matches per obs.	-	-	-	-	2	4	6	8
Observations	10,316	7,057	8,133	7,057	7,057	7,057	7,057	7,057
Science questionnaire								
Level								
PAE	0.022	0.029	0.026	0.035	-0.008	-0.016	-0.005	-0.002
	(0.026)	(0.023)	(0.025)	(0.023)	(0.034)	(0.029)	(0.028)	(0.027)
P25 of the entire sample								
PAE	0.006	0.006	0.001	0.002	0.019	0.019	0.018	0.015
	(0.012)	(0.010)	(0.012)	(0.010)	(0.014)	(0.013)	(0.012)	(0.012)
Controls	No	Yes	No	Yes	Yes	Yes	Yes	Yes
No. matches per obs.	-	-	-	-	2	4	6	8
Observations	10,877	7,411	8,550	7,411	7,411	7,411	7,411	7,411

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

Table 6: The impact of PAE on Rate decline by gender

	Boys				Girls			
	OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Complete questionnaire								
Level								
PAE	-0.048	-0.020	-0.011	-0.011	0.088***	0.088***	0.107***	0.093***
	(0.030)	(0.029)	(0.032)	(0.031)	(0.027)	(0.027)	(0.032)	(0.030)
P25 of the entire sample								
PAE	0.020	0.012	0.005	0.005	-0.040***	-0.042***	-0.046***	-0.045***
	(0.013)	(0.013)	(0.014)	(0.014)	(0.011)	(0.012)	(0.013)	(0.013)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,089	10,964	11,010	10,964	11,089	10,964	11,010	10,964
Maths questionnaire								
Level								
PAE	0.004	0.017	-0.013	0.009	0.036	0.005	0.038	0.016
	(0.029)	(0.028)	(0.033)	(0.029)	(0.029)	(0.026)	(0.032)	(0.028)
P25 of the entire sample								
PAE	0.011	0.007	0.018	0.012	-0.032**	-0.025**	-0.043***	-0.038***
	(0.013)	(0.013)	(0.014)	(0.014)	(0.012)	(0.012)	(0.014)	(0.013)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	10,570	10,437	10,487	10,437	10,570	10,437	10,487	10,437
Reading questionnaire								
Level								
PAE	0.001	0.024	0.051	0.044	-0.050	-0.014	-0.032	-0.023
	(0.038)	(0.035)	(0.041)	(0.037)	(0.033)	(0.034)	(0.036)	(0.034)
P25 of the entire sample								
PAE	0.001	-0.007	-0.019	-0.016	0.028*	0.012	0.022	0.014
	(0.015)	(0.015)	(0.017)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	10,316	7,057	8,133	7,057	10,316	7,057	8,133	7,057
Science questionnaire								
Level								
PAE	0.036	0.037	0.064*	0.055	0.008	0.021	-0.014	0.018
	(0.033)	(0.033)	(0.036)	(0.034)	(0.031)	(0.029)	(0.031)	(0.029)
P25 of the entire sample								
PAE	0.001	0.008	-0.022	-0.009	0.010	0.003	0.025	0.013
	(0.015)	(0.015)	(0.016)	(0.014)	(0.014)	(0.014)	(0.016)	(0.015)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	10,877	7,411	8,550	7,411	10,877	7,411	8,550	7,411

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

Table 7: Summary Statistics

Variable	(1) Girls	(2) Boys	(3) P-value Diff. (1)-(2)	(4) P-score Controls
Individual variables				
Initial test score[a]	.587	.627	.000	yes
	(.276)	(.27)		
Migrant(=1)	.153	.147	.601	yes
	(.36)	(.354)		
Repeated once(=1)	.242	.301	.000	yes
	(.429)	(.459)		
Repeated more than once(=1)	.084	.128	.000	yes
	(.278)	(.334)		
Attended kindergarden(=1)	.844	.816	.029	yes
	(.363)	(.387)		
Socioeconomic variables				
Index of educ possession	.095	-.013	.000	yes
	(.864)	(.907)		
Mother highly educated(=1)	.299	.304	.735	yes
	(.458)	(.46)		
Father highly educated(=1)	.278	.316	.011	no
	(.448)	(.465)		
School variables				
Student-Teacher Ratio	9.277	9.295	.791	yes
	(2.015)	(2.08)		
ESCS	-.377	-.371	.852	no
	(.976)	(.968)		
ESCS in high quartile(=1)	.146	.151	.652	yes
	(.353)	(.358)		
Prop. of dropout	.115	.116	.753	yes
	(.110)	(.112)		
Prob. of dropout in	.322	.299	.137	yes
high quartile(=1)	(.467)	(.458)		
School size	624.581	624.968	.966	yes
	(274.553)	(280.856)		
Prop. of migrants (school)	.15	.15	.945	no
	(.143)	(.145)		
Parental pressure	.396	.375	.188	no
on teachers(=1)	(.489)	(.484)		
School climate-teacher	.701	.684	.276	yes
	(.458)	(.465)		
Principal enhance	.22	.211	.535	no
school's reputation(=1)	(.414)	(.408)		
Rural(=1)	.411	.4	.484	no
	(.492)	(.49)		
Observations	1,829	1,831		

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.

Table 8: The impact of PAE on Rate decline (Subgroups)

	Boys				Girls				Overall			
	OLS		IPWE		OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Schools with many immigrants and repeaters												
	Complete questionnaire											
	Level											
PAE	-0.040	-0.006	0.005	0.006	0.077***	0.087***	0.082**	0.076**	0.018	0.041*	0.044*	0.041*
	(0.031)	(0.029)	(0.034)	(0.032)	(0.029)	(0.029)	(0.035)	(0.033)	(0.022)	(0.023)	(0.025)	(0.024)
	P25 of the entire sample											
PAE	0.015	0.004	-0.002	-0.002	-0.037***	-0.044***	-0.039***	-0.040***	-0.011	-0.020**	-0.020**	-0.021**
	(0.014)	(0.014)	(0.015)	(0.015)	(0.012)	(0.013)	(0.014)	(0.014)	(0.009)	(0.010)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	10,438	10,320	10,359	10,320	10,438	10,320	10,359	10,320	10,438	10,320	10,359	10,320
Non-educated families												
	Level											
PAE	-0.058	-0.040	-0.001	-0.010	0.058*	0.056*	0.067	0.059	0.001	0.009	0.034	0.025
	(0.037)	(0.036)	(0.043)	(0.042)	(0.035)	(0.032)	(0.043)	(0.039)	(0.027)	(0.026)	(0.031)	(0.030)
	P25 of the entire sample											
PAE	0.023	0.014	-0.001	0.002	-0.032***	-0.038***	-0.041**	-0.042**	-0.005	-0.013	-0.021*	-0.021*
	(0.017)	(0.017)	(0.019)	(0.019)	(0.014)	(0.014)	(0.017)	(0.017)	(0.011)	(0.011)	(0.012)	(0.012)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	9,537	9,417	6,752	6,711	9,537	9,417	6,752	6,711	9,537	9,417	6,752	6,711

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

Table 9: The impact of PAE on Rate decline at the School level

	OLS		IPWE	
	(1)	(2)	(3)	(4)
	Complete questionnaire			
	Level			
PAE	0.019	0.038*	0.043*	0.038*
	(0.021)	(0.022)	(0.023)	(0.021)
	P25 of the entire sample			
PAE	-0.010	-0.015	-0.018*	-0.017*
	(0.009)	(0.009)	(0.010)	(0.009)
Controls	No	Yes	No	Yes
Observations	395	395	395	395

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

Figures

Figure 1: Decline in Performance

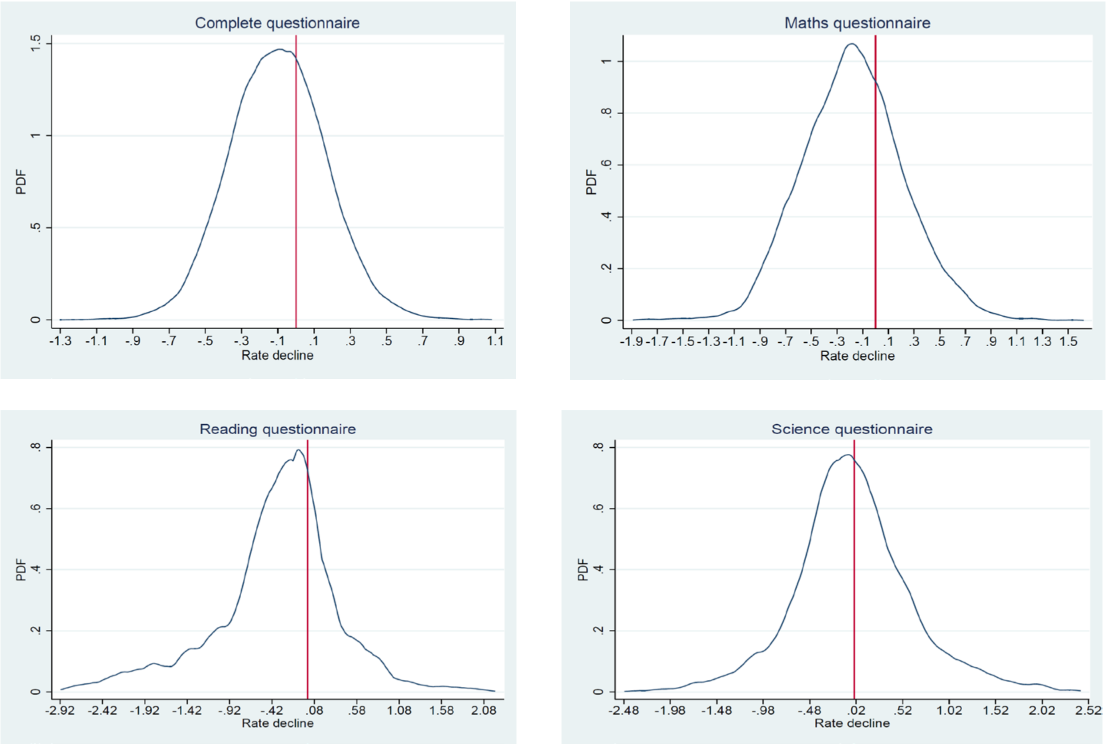
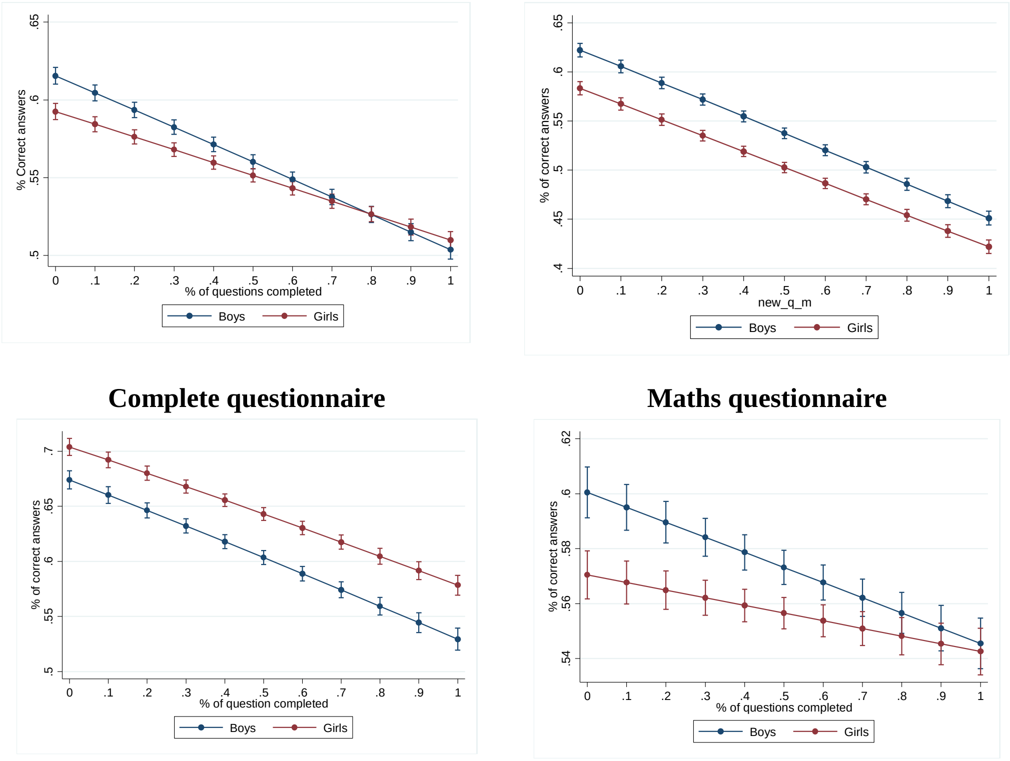


Figure 2: Decline in Performance: The gender gap



Complete questionnaire

Maths questionnaire

Reading questionnaire

Science questionnaire

Figure 3: Propensity Score Support

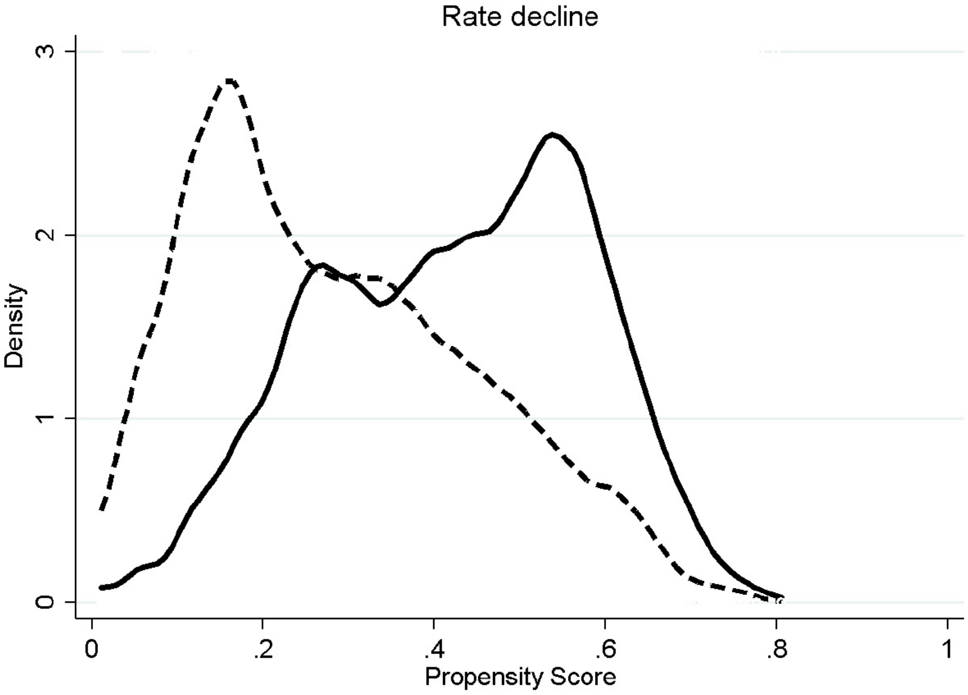
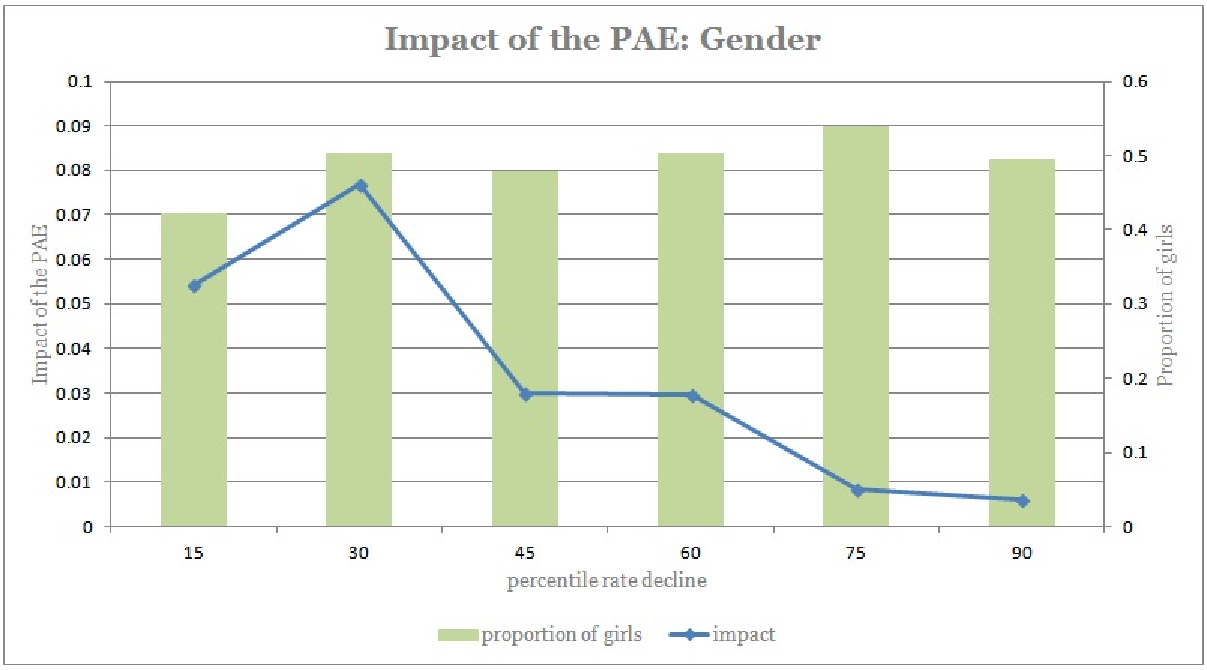


Figure 4: Impact of the PAE: Gender



Supplementary Material to:

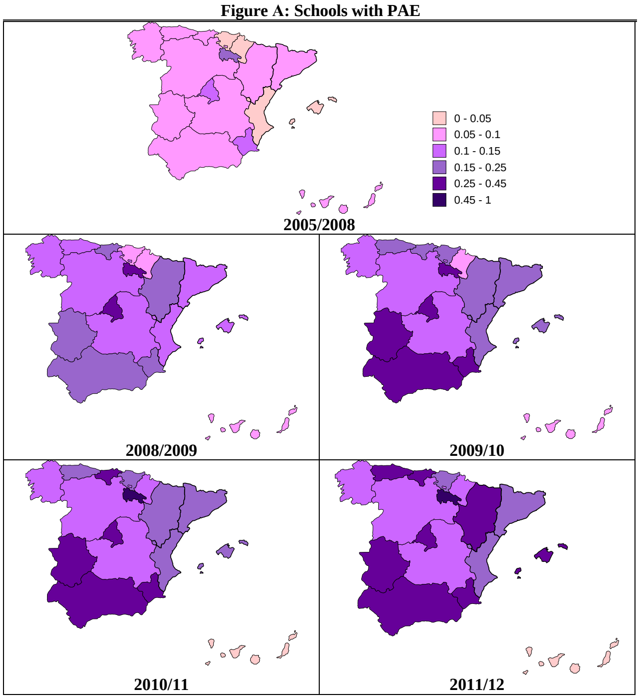
Test performance and Remedial Education: Good news for girls

Marianna Battaglia and Marisa Hidalgo-Hidalgo

A: The remedial program

In this Section we provide additional details on the PAE program. Figure A displays the percentage of public secondary schools in which the PAE was implemented in each region during the full period that the program was implemented, that is, from the 2005/06 until the 2011/12 academic year.

Figure A: Schools with PAE



Note: Proportion of schools with the PAE over total number of public schools. Source: García-Pérez e Hidalgo-Hidalgo (2017), INEE (Instituto Nacional de Evaluación Educativa) and Ministerio de Educación (2016)

Table A below shows the number of PAE secondary schools and participation year together with the number of schools that participated in the 2012 PISA program per region.

Table A: Schools with PAE in PISA 2012

		2005/2006		2006/2007		2007/2008		2008/2009		2009/2010		2010/2011		2011/2012	
	PISA12	PAE	PISA	PAE	PISA	PAE	PISA	PAE	PISA	PAE	PISA	PAE	PISA	PAE	PISA
Andalusia	52	37	4	72	3	161	9	200	11	320	16	350	16	400	7
Aragon	51	4	1	7	3	15	3	19	3	28	6	31	8	50	16
Asturias	56	3	2	5	2	11	3	11	5	11	6	11	6	11	6
Balearic Islands	54	0	0	0	0	0	0	10	6	15	10	15	14	26	16
Cantabria	54	2	2	4	4	8	7	10	9	10	5	18	6	19	6
Castile and Leon	55	8	3	15	5	33	5	36	6	36	7	36	7	36	7
Catalonia	51	20	0	36	0	71	4	71	5	92	4	92	4	92	4
Extremadura	53	6	2	11	2	23	4	23	4	37	8	50	11	54	15
Galicia	56	10	1	19	2	40	8	40	8	45	4	45	8	49	10
La Rioja	54	1	1	5	5	10	9	13	10	12	15	15	19	17	19
Madrid	51	11	1	26	2	78	6	100	9	109	11	114	11	126	14
Region of Murcia	52	6	1	11	5	26	11	28	10	39	14	51	19	51	18
Navarre	51	1	0	3	1	6	3	6	3	7	2	8	2	10	2
Basque Country	174	0	0	4	3	11	2	13	3	30	13	42	20	50	23
Rest	38	40	2	71	2	94	2	112	5	117	9	106	3	111	2
Total	902	149	20	289	39	587	76	692	97	908	130	984	154	1102	165

Note: *Rest* refers to those regions without enlarged sample in PISA 2012 (Canary Islands, Castilla-La Mancha, Ceuta, Melilla and Valencian Community).
Source: INEE (Instituto Nacional de Evaluación Educativa) and PISA 2012

The first column of Table A shows the number of secondary schools that participated in the PISA 2012 assessment per region. We use PISA 2012 test for the regions with enlarged samples: Andalusia, Aragon, Asturias, Balearic Islands, Cantabria, Castile and Leon, Catalonia, Extremadura, Galicia, La Rioja, Madrid, Region of Murcia, Navarre, and Basque Country. The table also shows the number of secondary schools where the PAE was implemented in a particular academic year, regardless of whether it was also implemented in other academic years, and the number of schools with PAE that also participated in the PISA 2012 assessment. As it can be observed, more than 10% of the schools where the PAE was implemented during 2011/12 were also evaluated in PISA 2012.

B: Alternative definitions of correct answer and difficulty

Table A: The impact of PAE on Rate decline - Partially corrected answers

	Boys				Girls				Overall			
	OLS		IPWE		OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
All sample												
							Level					
PAE	-0.045	-0.018	-0.014	-0.013	0.084***	0.083***	0.106**	0.090**	0.019	0.032	0.046*	0.039
	(0.030)	(0.030)	(0.032)	(0.031)	(0.028)	(0.028)	(0.033)	(0.031)	(0.022)	(0.023)	(0.024)	(0.024)
							P25 of the entire sample					
PAE	0.019	0.011	0.005	0.005	-0.046***	-0.048***	-0.053***	-0.051***	-0.013	-0.018*	-0.024**	-0.023**
	(0.014)	(0.014)	(0.014)	(0.014)	(0.012)	(0.012)	(0.013)	(0.013)	(0.009)	(0.010)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,088	10,963	11,009	10,963	11,088	10,963	11,009	10,963	11,088	10,963	11,009	10,963
Schools with many immigrants and repeaters												
							Level					
PAE	-0.040	-0.006	-0.002	0.000	0.071**	0.079***	0.080**	0.071**	0.015	0.037	0.039	0.036
	(0.031)	(0.030)	(0.034)	(0.032)	(0.030)	(0.030)	(0.037)	(0.035)	(0.022)	(0.023)	(0.026)	(0.025)
							P25 of the entire sample					
PAE	0.018	0.006	0.001	0.000	-0.042***	-0.048***	-0.045***	-0.045***	-0.012	-0.021**	-0.022**	-0.023**
	(0.015)	(0.015)	(0.015)	(0.015)	(0.013)	(0.013)	(0.015)	(0.014)	(0.010)	(0.010)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	10,437	10,319	10,358	10,319	10,437	10,319	10,358	10,319	10,437	10,319	10,358	10,319
Non-educated families												
							Level					
PAE	-0.055	-0.036	-0.007	-0.012	0.057	0.053	0.067	0.057	0.002	0.009	0.031	0.023
	(0.037)	(0.036)	(0.043)	(0.042)	(0.035)	(0.032)	(0.044)	(0.040)	(0.027)	(0.027)	(0.031)	(0.030)
							P25 of the entire sample					
PAE	0.021	0.013	0.003	0.003	-0.040***	-0.046***	-0.047**	-0.049**	-0.010	-0.017	-0.023*	-0.023*
	(0.017)	(0.017)	(0.018)	(0.018)	(0.014)	(0.013)	(0.016)	(0.016)	(0.011)	(0.011)	(0.012)	(0.012)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	9,536	9,416	6,751	6,710	9,536	9,416	6,751	6,710	9,536	9,416	6,751	6,710

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

Table B: The impact of PAE on Rate decline - Different measures of difficulty

	Boys				Girls				Overall			
	OLS		IPWE		OLS		IPWE		OLS		IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
All sample - Difficulty equal to 1 if the answer is an open question												
						Level						
PAE	-0.043	-0.020	-0.018	-0.016	0.071**	0.066**	0.082***	0.069**	0.013	0.023	0.033	0.026
	(0.030)	(0.029)	(0.032)	(0.032)	(0.028)	(0.027)	(0.031)	(0.030)	(0.022)	(0.023)	(0.024)	(0.023)
						P25 of the entire sample						
PAE	0.025*	0.018	0.013	0.012	-0.031***	-0.032***	-0.035***	-0.034***	-0.003	-0.007	-0.011	-0.011
	(0.014)	(0.014)	(0.015)	(0.015)	(0.011)	(0.012)	(0.013)	(0.013)	(0.009)	(0.010)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,075	10,955	11,001	10,955	11,075	10,955	11,001	10,955	11,075	10,955	11,001	10,955
All sample - Difficulty as percentage of students who correctly answer to the question												
						Level						
PAE	-0.055*	-0.029	-0.016	-0.019	0.087***	0.094***	0.110**	0.101***	0.016	0.033	0.047*	0.041*
	(0.031)	(0.031)	(0.034)	(0.033)	(0.027)	(0.027)	(0.030)	(0.029)	(0.022)	(0.023)	(0.024)	(0.023)
						P25 of the entire sample						
PAE	0.027*	0.017	0.007	0.009	-0.041***	-0.047***	-0.051***	-0.050***	-0.007	-0.015	-0.022**	-0.021**
	(0.014)	(0.014)	(0.015)	(0.015)	(0.011)	(0.012)	(0.013)	(0.013)	(0.009)	(0.010)	(0.010)	(0.010)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,091	10,963	11,011	10,963	11,091	10,963	11,011	10,963	11,091	10,963	11,011	10,963

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

C: Reading or Maths first

Table A: The impact of PAE on Rate decline - Clusters order

	Boys		Girls		Overall		
	OLS	IPWE	OLS	IPWE	OLS	IPWE	
	(1)	(2)	(3)	(4)	(5)	(6)	Observations
All sample - Complete questionnaire							
				Level			
Reading after Maths	-0.009	-0.005	0.105**	0.099**	0.050	0.048	4,238
	(0.047)	(0.050)	(0.044)	(0.049)	(0.032)	(0.031)	
Maths after Reading	-0.031	-0.019	0.076**	0.084**	0.022	0.032	6,726
	(0.035)	(0.038)	(0.032)	(0.037)	(0.025)	(0.024)	
Chi2 test (p-value)	0.2147	0.0097	0.4162	0.0926			
				P25 of the entire sample			
Reading after Maths	0.021	0.021	-0.033*	-0.033*	-0.007	-0.007	4,238
	(0.022)	(0.022)	(0.020)	(0.020)	(0.015)	(0.014)	
Maths after Reading	0.007	0.007	-0.046***	-0.046***	-0.020*	-0.026**	6,726
	(0.016)	(0.016)	(0.015)	(0.015)	(0.011)	(0.011)	
Chi2 test (p-value)	0.5817	0.0105	0.9996	0.8077			

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

As reported in Table A the order of the subject does not show to be relevant for the rate of decline in the complete questionnaire. We do 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).

D: Item reached

Table A: Students' outcomes: non-cognitive skills (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)
Complete questionnaire	.973	.971	.973	.967	.974	.97	.976	.971	.973	.97	.975	.969
	(.072)	(.079)	(.068)	(.077)	(.063)	(.073)	(.057)	(.064)	(.068)	(.076)	(.063)	(.071)
First quartile (P25)	.286	.292	.282	.32	.297	.307	.293	.33	.292	.3	.288	.325
Observations	5,429	1,823	3,606	3,606	5,581	1,827	3,754	3,754	11,010	3,650	7,360	7,360

Standard deviations in parentheses.

Table B: 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)
All sample												
	Level											
PAE	-0.001	0.003	0.004	0.005*	-0.005*	-0.000	-0.001	0.000	-0.003	0.001	0.001	0.003
	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.013)	(0.002)	(0.003)	(0.002)
	P25 of the entire sample											
PAE	0.008	-0.019	-0.028	-0.024	0.013	-0.016	-0.023	-0.023	0.010	-0.018	-0.025*	-0.023*
	(0.016)	(0.015)	(0.019)	(0.017)	(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	11,010	10,964	11,089	10,964	11,010	10,964	11,089	10,964	11,010	10,964

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

E: School characteristics

Table A: 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]	.63	.624	.632	.588	.637	.694	yes
	(.143)	(.141)	(.145)		(.138)		
Girl(=1)	.501	.497	.503	.664	.504	.654	yes
	(.123)	(.123)	(.123)		(.091)		
Migrant(=1)	.125	.179	.099	.000	.166	.359	yes
	(.17)	(.205)	(.143)		(.184)		
Repeated once(=1)	.247	.284	.229	.001	.263	.477	yes
	(.144)	(.157)	(.133)		(.103)		
Repeated more than once(=1)	.104	.123	.095	.058	.111	.624	yes
	(.138)	(.144)	(.135)		(.093)		
Attended kindergarden(=1)	.834	.825	.838	.361	.829	.941	yes
	(.138)	(.141)	(.136)		(.118)		
Socioeconomic variables							
Index of educ possession	.036	.024	.042	.574	.071	.249	yes
	(.3)	(.296)	(.302)		(.232)		
Mother highly educated(=1)	.327	.287	.347	.000	.302	.963	yes
	(.17)	(.143)	(.179)		(.154)		
School variables							
Student-Teacher Ratio	9.441	9.108	9.602	.372	9.327	.865	yes
	(7.048)	(2.158)	(8.457)		(2.545)		
ESCS in high quartile(=1)	.248	.14	.301	.000	.159	.782	yes
	(.432)	(.348)	(.459)		(.367)		
Prop. of dropout	.103	.127	.092	.005	.121	.675	yes
	(.114)	(.116)	(.118)		(.119)		
Prob. of dropout in	.23	.302	.195	.025	.331	.682	yes
high quartile(=1)	(.422)	(.461)	(.397)		(.472)		
School size	581.171	599.875	572.1	.404	628.508	.900	yes
	(325.934)	(293.509)	(340.709)		(272.071)		
School climate-teacher	.554	.674	.496	.001	.717	.620	yes
	(.498)	(.47)	(.501)		(.451)		
Observations	395	129	266		266		

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.

F: Self-assessed measures

We examine here the impact of the program on students' self-assessed measures. In particular 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. Discipline is measured by the way students behave in class (disciplinary climate). 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 belong 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 below.

Overall, we observe that discline improves, especially for boys and in schools with many immigrants and repeaters, and perception of learning at school slightly increases for girls, especially in non-educated families (Tables B and C).

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

	(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	.252	.231	.274	.241	.255	.234	.284	.239	.254	.232	.279
	(.426)	(.434)	(.422)	(.446)	(.428)	(.436)	(.423)	(.451)	(.427)	(.435)	(.422)	(.449)
Observations	5,413	1,812	3,601	3,601	5,570	1,818	3,752	3,752	10,983	3,630	7,353	7,353
Truancy(=1)	.366	.385	.357	.376	.364	.384	.354	.38	.365	.384	.355	.378
	(.487)	(.487)	(.479)	(.486)	(.481)	(.486)	(.478)	(.485)	(.481)	(.486)	(.479)	(.485)
Observations	5,381	1,803	3,578	3,578	5,540	1,804	3,736	3,736	10,921	3,607	7,314	7,314
Discipline												
Bad climate(=1)	.39	.373	.398	.408	.357	.342	.365	.362	.373	.357	.381	.385
	(.488)	(.484)	(.49)	(.492)	(.479)	(.474)	(.481)	(.481)	(.484)	(.479)	(.486)	(.487)
Observations	5,448	1,831	3,617	3,617	5,587	1,829	3,758	3,758	11,035	3,660	7,375	7,375
Self-confidence(=1)	.287	.28	.291	.293	.26	.252	.265	.259	.273	.265	.277	.276
	(.452)	(.449)	(.454)	(.455)	(.439)	(.433)	(.441)	(.438)	(.446)	(.441)	(.448)	(.447)
Observations	5,448	1,831	3,617	3,617	5,587	1,829	3,758	3,758	11,035	3,660	7,375	7,375
Sense of belonging(=1)	.952	.957	.949	.95	.973	.974	.973	.968	.963	.966	.962	.96
	(.214)	(.204)	(.219)	(.217)	(.162)	(.16)	(.163)	(.170)	(.188)	(.182)	(.191)	(.197)
Observations	2,992	1,015	1,977	1,977	3,382	1,106	2,276	2,276	6,374	2,121	4,253	4,253
Perception of learning	.586	.588	.586	.586	.641	.646	.638	.633	.614	.617	.612	.609
at school(=1)	(.493)	(.492)	(.493)	(.493)	(.48)	(.478)	(.481)	(.482)	(.487)	(.486)	(.487)	(.488)
Observations	5,448	1,831	3,617	3,617	5,587	1,829	3,758	3,758	11,035	3,660	7,375	7,375

Standard deviations in parentheses.

Table B: The impact of PAE on Non Cognitive Self-assessed 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	0.0206	-0.0163	-0.0221	-0.0151	0.0209	-0.0199	-0.0291	-0.0283	0.0662	-0.0633	-0.0779	-0.0702
	(0.0167)	(0.0151)	(0.0196)	(0.0177)	(0.0187)	(0.0160)	(0.0203)	(0.0176)	(0.0497)	(0.0471)	(0.0536)	(0.0488)
Truancy												
PAE	0.0263	0.0116	0.0086	0.0139	0.0290	0.0108	0.0035	0.0046	0.0733*	0.0307	0.0160	0.0251
	(0.0175)	(0.0170)	(0.0192)	(0.0180)	(0.0201)	(0.0191)	(0.0223)	(0.0200)	(0.0422)	(0.0419)	(0.0454)	(0.0412)
Discipline												
Bad Climate												
PAE	-0.0235	-0.0269*	-0.0348**	-0.0329**	-0.0204	-0.0245	-0.0200	-0.0191	-0.0575*	-0.0687**	-0.0728**	-0.0693**
	(0.0145)	(0.0151)	(0.0165)	(0.0162)	(0.0158)	(0.0160)	(0.0168)	(0.0167)	(0.0304)	(0.0319)	(0.0331)	(0.0328)
Self-confidence												
PAE	-0.0098	-0.0093	-0.0131	-0.0137	-0.0137	-0.0119	-0.0088	-0.0075	-0.0348	-0.0324	-0.0332	-0.0324
	(0.0122)	(0.0125)	(0.0132)	(0.0131)	(0.0125)	(0.0131)	(0.0148)	(0.0143)	(0.0265)	(0.0285)	(0.0304)	(0.0293)
Sense of belonging												
PAE	0.0069	0.0088	0.0064	0.0049	0.0014	0.0034	0.0057	0.0064	0.0470	0.0777	0.0737	0.0728
	(0.0084)	(0.0087)	(0.0090)	(0.0089)	(0.0061)	(0.0062)	(0.0075)	(0.0071)	(0.0659)	(0.0726)	(0.0748)	(0.0724)
Perception of learning at school												
PAE	0.0042	0.0116	0.0023	0.0073	0.0115	0.0140	0.0133	0.0147	0.0194	0.0338**	0.0205	0.0289
	(0.0127)	(0.0124)	(0.0142)	(0.0135)	(0.0116)	(0.0116)	(0.0127)	(0.0126)	(0.0173)	(0.0172)	(0.0202)	(0.0186)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	11,105	10,998	11,035	10,998	11,105	10,998	11,035	10,998	11,105	10,998	11,035	10,998

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

Table C: The impact of PAE on Non Cognitive Self-assessed Outcomes (Subgroups)

	Schools with many immigrants and repeaters						Non-educated families					
	Boys		Girls		Overall		Boys		Girls		Overall	
	OLS	IPWE	OLS	IPWE	OLS	IPWE	OLS	IPWE	OLS	IPWE	OLS	IPWE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Motivation												
						Absenteism						
PAE	-0.0155	-0.0125	-0.0049	-0.0120	-0.0331	-0.0378	-0.0056	0.0058	-0.0127	-0.0299	-0.0297	-0.0385
	(0.0183)	(0.0210)	(0.0196)	(0.0208)	(0.0534)	(0.0551)	(0.0210)	(0.0237)	(0.0217)	(0.0222)	(0.0591)	(0.0603)
						Truancy						
PAE	-0.0016	-0.0002	-0.0030	-0.0107	-0.0064	-0.0147	0.0180	0.0137	0.0105	0.0015	0.0394	0.0205
	(0.0196)	(0.0212)	(0.0217)	(0.0231)	(0.0477)	(0.0473)	(0.0218)	(0.0230)	(0.0213)	(0.0234)	(0.0490)	(0.0494)
Discipline												
						Bad Climate						
PAE	-0.0403**	-0.0490***	-0.0047	0.0025	-0.0598*	-0.0616*	0.0028	-0.0049	-0.0126	-0.0129	-0.0139	-0.0241
	(0.0171)	(0.0184)	(0.0179)	(0.0189)	(0.0358)	(0.0369)	(0.0196)	(0.0202)	(0.0204)	(0.0210)	(0.0398)	(0.0406)
						Self-confidence						
PAE	-0.0204	-0.0183	-0.0052	-0.0047	-0.0386	-0.0350	-0.0135	-0.0119	-0.0400**	-0.0302	-0.0876**	-0.0678
	(0.0144)	(0.0150)	(0.0151)	(0.0164)	(0.0328)	(0.0327)	(0.0188)	(0.0201)	(0.0172)	(0.0191)	(0.0430)	(0.0439)
						Sense of belonging						
PAE	0.0080	0.0009	-0.0004	0.0053	0.0440	0.0409	0.0089	0.0051	0.0010	0.0027	0.0658	0.0542
	(0.0102)	(0.0102)	(0.0074)	(0.0088)	(0.0807)	(0.0837)	(0.0107)	(0.0110)	(0.0077)	(0.0087)	(0.0981)	(0.0962)
						Perception of learning at school						
PAE	0.0011	-0.0065	0.0166	0.0257*	0.0235	0.0254	0.0222	0.0140	0.0377**	0.0353*	0.0801***	0.0659**
	(0.0143)	(0.0157)	(0.0133)	(0.0148)	(0.0198)	(0.0217)	(0.0190)	(0.0205)	(0.0176)	(0.0181)	(0.0310)	(0.0325)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,656	7,656	7,656	7,656	7,656	7,656	5,688	5,688	5,688	5,688	5,688	5,688

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