Borradores de ECONOMÍA

Acing the Test: Educational Effects of the SaberEs Test Preparation Program in Colombia

By: Christian Posso
Estefanía Saravia
Pablo Uribe

No. 1237
2023

BANCO DE LA REPÚBLICA
COLOMBIA

Acing the Test: Educational Effects of the *SaberEs* Test Preparation Program in Colombia

Christian Posso[1], Estefanía Saravia[2], and Pablo Uribe[*3]

[1]Banco de la República (cpossosu@banrep.gov.co)
[2]Icfes and University of California, Los Angeles (esaravia@ucla.edu)
[3]Universidad Eafit (puribeb3@eafit.edu.co)

Abstract

Education in Colombia and Latin America is characterized by significant gaps in the quality of education as measured by standardized test scores. This paper assesses the impact of a Colombian program called *SaberEs*, which strengthens preparation for standardized cognitive tests such as the high school exit exam in Colombia (*Saber 11*) for low socioeconomic status students. The program provides competency-based training sessions to develop skills for analyzing and solving specific types of questions within school hours. Our difference-in-differences estimates show that *SaberEs* increased *Saber 11* scores by 2.22 ranks (or 0.074 standard deviations), which implies that the socioeconomic achievement gap was reduced by 23% regarding the control schools. Also, students affected by the program experienced a significant increase in access to tertiary education and merit-based scholarships in Colombia.

Keywords: Test preparation programs; standardized tests; higher education; financial aid and scholarships; education in Latin America.

JEL Classification: A21, D04, I21, I24, I28.

[*]We thank Naomi Calle for excellent research assistance. We thank Leonardo Bonilla, Mónica Hernandez, Camilo Acosta, Nicolás Mancera, Natalia Cantet, Brian Feld and seminar participants at Universidad Eafit, Universidad de los Andes, International Seminar of Icfes, and Universidad Icesi. We also thank the *Instituto Colombiano para la Evaluación de la Educación* - Icfes, the *Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior* – ICETEX, Sapiencia, and the Ministry of Education for providing access to the data and insightful discussions.

Aprobando el examen: Efectos educativos del programa SaberEs para preparación de pruebas en Colombia

Christian Posso, Estefanía Saravia, and Pablo Uribe

Resumen

La educación en Colombia y América Latina se caracteriza por brechas significativas en la calidad de la educación medida por los resultados de pruebas estandarizadas. Este trabajo evalúa el impacto de un programa colombiano llamado SaberEs, que fortalece la preparación para pruebas cognitivas estandarizadas como el examen de egreso de bachillerato en Colombia (Saber 11) para estudiantes de bajo nivel socioeconómico. El programa ofrece sesiones de entrenamiento basadas en competencias para desarrollar habilidades para analizar y resolver tipos específicos de preguntas dentro del horario escolar. Nuestras estimaciones de diferencia en diferencias muestran que SaberEs aumentó los puntajes de Saber 11 en 2,22 rankings (o 0,074 desviaciones estándar), lo que implica que la brecha de rendimiento socioeconómico se redujo en un 23 % con respecto a las escuelas de control. Asimismo, los estudiantes afectados por el programa experimentaron un aumento significativo en el acceso a la educación terciaria y a las becas por mérito en Colombia.

Palabras clave: Programas de preparación de exámenes; exámenes estandarizados; enseñanza superior; financiación y becas; educación en América Latina.

Clasificación JEL: A21, D04, I21, I24, I28.

1 Introduction

Over the last five decades, post-primary education has expanded dramatically in most low- and middle-income countries (Ferreyra, Avitabile, Paz, Botero, & Urzúa, 2017; World Bank, 2018). Although progress has been made in expanding access to education, significant gaps exist in achievement and quality of education, as measured by standardized test scores (Angrist & Lavy, 2009; Gneezy et al., 2019).

Performance on standardized tests matters. They hold the educational system accountable, hence their extensive use in measuring educational quality within and across countries. In particular, a major interest is seen in policies that improve the results of high-stakes tests at the end of secondary education. High-stakes tests can affect the transition to higher education and labor market outcomes (Angrist & Lavy, 2009; Bond, Bulman, Li, & Smith, 2018; Brunello & Kiss, 2022). In some contexts, the scores in such tests determine eligibility for financial aid and scholarships (Bernal & Penney, 2019; Bruce & Carruthers, 2014; Gurantz & Odle, 2020; Londoño-Vélez, Rodríguez, & Sánchez, 2020; Melguizo, Sanchez, & Velasco, 2016).[1]

Little is known about the inequalities that exist in the performance on high-stakes exams, which are associated with the testing process itself (Duquennois, 2022; Goodman, Gurantz, & Smith, 2020). There is evidence that a significant source of inequality in this context is due to the investments that socioeconomically advantaged families do in test preparation (Dobbie & Fryer, 2011; Duquennois, 2022; Page & Scott-Clayton, 2016). As a result, students with similar skills but different socioeconomic backgrounds may perform differently. This paper explores a policy that may be effective in closing such gaps.

This paper aims to provide new evidence on the effectiveness of standardized test preparation programs by analyzing *SaberEs*[2], an extracurricular[3] test preparation program aimed at economically disadvantaged students implemented in the city of Medellin-Colombia in 2016.[4] Preparation for these types of tests is expensive, particularly in the Latin American context. In Colombia, a regular test preparation program may cost about three times the monetary poverty line per capita. This program was offered for free.

The objective of the program was to strengthen preparation for standardized cognitive tests. In particular, for the high school exit exam called *Saber 11*.[5] The policy aimed to improve access to tertiary education for students graduating from public schools and enable them to compete for financial aid and scholarships that typically use the *Saber 11* score as a selection mechanism.[6]

The program's main components were as follows: first, providing competency-based

[1] They are also informative of the student's cognitive abilities, which can be predictive of college performance (Cohn, Cohn, Balch, & Bradley Jr, 2004).

[2] Official Program's website: https://www.medellin.edu.co/saberes/

[3] It was extracurricular but within school hours.

[4] Several cross-country studies have shown that students with higher socioeconomic status participate more in theses type of programs, especially in countries where testing has higher stakes (Byun, Chung, & Baker, 2018; Zwier, Geven, & van de Werfhorst, 2020).

[5] The exam is similar to the SAT and ACT in the US. Its scores are required for the admission processes of higher education institutions.

[6] For instance, access to the most important merit-based scholarship in Colombia requires a student to be above the 90th percentile of the *Saber 11* distribution (Londoño-Vélez et al., 2020). Likewise, access to financial aid through national and local educational credit agencies has score requisites depending on the credit's conditions.

training sessions to develop skills for analyzing and solving specific questions. To achieve this purpose, the program implemented a cascade system in which teachers received initial training and then transferred the information to students during regular class hours. Second, the program established the use of simulation exams that functioned as a familiarization with the real test, *Saber 11*.[7] Third, the program offered vocational counseling to senior students. Interestingly, the program contracted with the same firms that typically provide extracurricular test preparation to some of the city's elite private institutions to provide these services. During the first wave of the program, that took place between June and July of 2016, *Saber 11*'s test rankings for the targeted public institutions improved in Medellin. It was not long before the press and politicians anecdotally claimed that the program had a positive impact, significantly improving the test scores of public schools in 2016.[8]

In this paper, we address three questions. First, Does the *SaberEs* program affect student learning gains measured by *Saber 11* scores? Although there is evidence that this type of program has positive effects on students with higher socioeconomic status, evidence on the causal relationship between this type of program and test results for disadvantaged students is lacking.[9] Second, Does the program affect access to tertiary education? To answer this question, we linked all students in our setting to administrative records on access to tertiary education, public financial aid at the national and local level, and Colombia's largest merit-based scholarship, *Ser Pilo Paga* (Londoño-Vélez et al., 2020). Using these administrative records, we could track students up to three years after the beginning of the program. Third, what mechanisms made *SaberEs* successful in increasing access to tertiary education programs? We argue that the program may have mainly affected access to higher education through the accumulation of specific human capital. That is, skills to analyze and solve the questions of high-stakes standardized exams. Another potential channel is higher access to merit-based scholarships.

This paper begins in section 2 by providing a detailed description of the educational system in Colombia and the *SaberEs* program. We then, in section 3, provide a comprehensive description of the administrative records used in the paper. Additionally, in section 4, we describe how the timing and implementation of the program are key to our empirical strategy. The main identification problem in this context is the selection of schools into the program. Taking advantage of the lack of universal coverage in the first year of the policy and its timing, we use difference-in-differences (DiD), and staggered event study approaches to estimate the causal effects of the program.

During the program's first year, the allocated budget allowed the test preparation firms to conduct a vocational guidance test with all public-school students. However, the test preparation component could only be given to a limited number of schools. The schools were selected primarily based on their previous *Saber 11* test scores,[10] targeting the

[7]The first two strategies were accompanied by additional investments to enable the implementation of both teacher and student preparation, as well as simulation exams.

[8]https://bit.ly/3ovyYSW. In addition, the vocational guidance component was also recognized for providing students with sufficient information to make wiser decisions about their post-secondary education https://bit.ly/3PGQxv4

[9]In the particular case of Colombia, Gómez, Bernal, and Herrera (2020) finds some heterogeneous effects associated with preparation exams but the effects are concentrated on private institutions.

[10]Although most of the schools were covered by 2017, it was only up to 2019 that all public institutions in the city were covered.

schools at the bottom of the previous year's score distribution, so the selected institutions could not anticipate their inclusion into the program. The program was implemented in assigned schools from grades 8 through 11 (senior year in Colombia), although most of its resources were concentrated on senior students. Even though the assignment mechanism was not perfect (i.e., some schools at the bottom of the distribution were not treated), treated schools performed significantly lower in the pre-treatment period than the control group on average.

Furthermore, we alleviate the identification concerns by combining a different set of administrative records with new causal inference methods. As an initial step toward mitigating the selection problem, we added an extensive set of controls to our DiD specification.[11] We also use this additional set of covariates to further control in a non-linear and non-parametric fashion. In particular, we use the Outcome Regression (OR) method proposed by Heckman, Ichimura, and Todd (1997) and Inverse Probability Weighting (IPW) from Abadie (2005).[12] Moreover, we use Sant'Anna and Zhao (2020) doubly robust difference-in-differences (DR) regression to obtain a more efficient and reliable estimator. This method combines OR and IPW to obtain a robust estimate if at least one of the two models is correctly specified, allowing for greater flexibility.

In addition to the DiD methods described above, we also use a dynamic DiD to estimate the program's causal effect on students' achievement. Specifically, we use Two-Way Fixed Effects (TWFE) to estimate an event study with the dynamic effects of the policy. Furthermore, we took advantage of the fact that in 2017 a second wave of schools entered the program to estimate a staggered event study, in which we have multiple periods and two years of treatment. Given that a simple TWFE regression would potentially be biased due to the presence of heterogeneous effects (Borusyak & Jaravel, 2017; De Chaisemartin & d'Haultfoeuille, 2020),[13] we overcome this challenge by using the Callaway and Sant'Anna (2021) estimator, which estimates all group-time-specific effects. We also aggregate the event study estimates to a single coefficient following Callaway and Sant'Anna (2021) and Borusyak, Jaravel, and Spiess (2021). Finally, since all these methods are based on the parallel trends assumption (Roth, Sant'Anna, Bilinski, & Poe, 2022), we estimate the possible bias of the prior tests following Roth (2022) and perform a sensitivity analysis to check the robustness of the results to certain violations of the parallel trends assumption (Rambachan & Roth, 2023).

Our results are consistent across specifications. In section 5, we show that *SaberEs* significantly increased students' test scores. Our results suggest that this test-training program, on average, increased the students' rank on the *Saber 11* test between 2.22 and 2.96, depending on the specification.[14] These results represent a reduction in the rank's gap between treated and untreated students of 23%. The positive effect on the total score is primarily driven by the effects on math, science, and social studies scores. The program did not affect the results of the English test scores, which is to be expected given that the program was not designed to teach a second language and could have hardly

[11] These include school characteristics and self-reported socioeconomic characteristics such as household goods, parent's education, employment status and socioeconomic status.

[12] See also, Hájek (1971); Horvitz and Thompson (1952).

[13] This happens because the estimator is a weighted average of all 2x2 comparisons and therefore includes "forbidden comparisons" that could have negative weights and change the sign of the estimate (Goodman-Bacon, 2021).

[14] Alternatively, it led to a 0.07 standard deviations increase in the *Saber 11* standardized test scores.

done so in such a short time. *SaberEs* did not affect the bottom of the distribution. Using unconditional quantile regression (Firpo, Fortin, & Lemieux, 2009), we show that the effects are statistically significant only for students at the 45th percentile or above. Although the effects are stronger around the median, the program positively affects the upper part of the distribution, including percentiles above the *Ser Pilo Paga* (SPP) cut-off.[15] These results contribute to the literature on the effects of public test preparation policies on test scores in Colombia. The only exception is Gómez et al. (2020), who study several preparation programs using cross-sectional data from Colombia and propensity score matching methods.

In addition to these results, we show that *SaberEs* had a substantial effect on access to higher education, the program's main target. Our estimates suggest that the program led to a 3.7 percentage points increase in post-secondary enrollment on average. This corresponds to a relative effect of around 5% from a baseline value of 67.7%. Separating access by type of program and years after high school graduation, we find that results are positive across years and are mainly driven by access to short-cycle programs.[16] Interestingly, we also find an average positive effect on access to STEM programs of 2.1 percentage points, corresponding to a relative effect of 7.3% from a baseline level of 28.7%. Not only did students access short-cycle programs to a higher degree, but we also show that *SaberEs* increased graduation from these programs by 2.3 percentage points on average, relative to a baseline average of 15.2%.

Section 6 discusses several potential mechanisms by which the program affects access to higher education. Given that the program's main strategy was to develop skills in the analysis and solution of standardized tests, one potential mechanism is the accumulation of task-specific human capital that was especially helpful when students took high-stakes tests such as admission exams of tertiary education institutions. In Medellin, the largest public higher institutions, *Universidad de Antioquia* (UdeA), *Universidad Nacional*, and SENA,[17] require applicants to take standardized tests as part of their admission processes.[18] We look at the effects on access to these institutions versus institutions that do not require their own standardized exams.[19] Although we do not find effects on access to professional programs, we find that access to short-cycle programs in institutions with these tests increased significantly. On the contrary, we do not find any significant effects on higher education institutions that do not have admission exams.

These effects can result from task-specific human capital and/or motivational effects. Under motivational effects, if students are motivated to learn, you expect to see an improvement in their overall scores regardless of the relative importance of the exam. To test this idea, we use a low-stakes exam unique to the city of Medellin. Every year the

[15] SPP required students to score approximately above the 90th percentile in their cohort. In 2016, this corresponded to getting a total score in *Saber 11* of 342 or higher.

[16] These are technical and technological programs, which have a duration of 2-3 years, respectively.

[17] *Servicio Nacional de Aprendizaje* (SENA) is a public establishment ascribed to the Ministry of Labor that offers free training in short-cycle programs that contribute to the country's economic, scientific and social development.

[18] In these institutions, only students with the highest scores are admitted. The final scores of these tests are not publicly available. For that reason we focus exclusively on access.

[19] *Saber 11* is required in the admission processes of all higher education institutions in the country (Londoño-Vélez et al., 2020). However, aside from specific cases, it is not used to discriminate between applicants.

Secretary of Education carries out *Olimpiadas del Conocimiento*, a series of competitions between students in grade 5, and in grades 10 and 11. *Olimpiadas del Conocimiento* has several stages, but only the first one applies a standardized test similar to *Saber 11* to all students.[20] Using the same methods described above, we find no discernible impact of the program on *Olimpiadas del Conocimiento*'s scores, so we rule out the motivational effect. Similar to the literature in labor economics that shows that task-specific human capital is an important source of individual wage growth (Gathmann & Schönberg, 2010; Gibbons & Waldman, 2004), here we show that it is also associated with strong effects on access to tertiary education.

A second mechanism is through financial aid and scholarships. In Colombia, *Saber 11* is used as a selection mechanism to access the main sources of financial aid and scholarships. Like Bernal and Penney (2019); Londoño-Vélez et al. (2020); Melguizo et al. (2016), we test the effects of *SaberEs* on such outcomes. The program had no discernible impact on access to public financial aid.[21] Yet it significantly affected access to SPP, the most important merit-based scholarship in the country. Our results indicate that it increased access to this scholarship by 1.1 percentage points on average. This corresponds to a striking 23.7% relative effect from a baseline value of 4.64%. The SPP scholarship allows low-income students to enter tertiary education, especially those at the top of the distribution of skills.

2 The Colombian postsecondary education system

The Colombian secondary education system runs from sixth through eleventh grade and ends with the application of a mandatory standardized test called *Saber 11*. Upon graduation, students may decide to move on to higher education, to either a short-cycle or a professional program. Regardless of the type of program in which they are enrolled, students are required by law to take a standardized test at the end of their program (*Saber TyT* or *Saber Pro*),[22] to be able to graduate. After this, they can enter the job market or continue their studies in graduate education.

2.1 Saber 11

The *Saber 11* high school exit exam is a mandatory test similar to the SAT in the United States, and administered by the *Instituto Colombiano para el Fomento de la Educación Superior* (Icfes), the institution responsible for measuring the quality of education through standardized testing in Colombia. *Saber 11* is a mandatory test with compliance rates

[20] In 2016, there were over 83.000 students in grades 5, 10 and 11 participating in the first stage between both categories (grade 5, and grades 10-11), with the 3.500 best students moving on to the second stage. From here, 25 students per category were selected to participate in the semifinals and only 5 of them moved on to the finals, with a single winner within each group. Only the ten finalists received the top prizes (trip to the USA for the grade 5 category, and a full university scholarship for the grade 10-11 finalists).

[21] We had access to the main public sources of financial aid in Colombia. First, the Colombian Institute of Educational Credit and Technical Studies Abroad (ICETEX by its acronym in Spanish) controls most of the educational credit market at the national level. Second, Medellin's Higher Education Agency (Sapiencia) which is especially relevant in the city.

[22] *Saber TyT* is taken by short-cycle students while *Saber Pro* is taken by professional students.

above 90% and is a good indicator of a student's cognitive skills (Bernal & Penney, 2019). An average of 500,000 students take it each year, in either spring or fall depending on the school's academic calendar.[23] All public schools are on calendar A, while elite private schools are usually on calendar B. The vast majority of the country's students are on calendar A and therefore, take the *Saber 11* in fall (August, middle of their school year).

The test has had several structural changes over the years, the most recent one happening in 2014. Prior to that year, the exam was divided into 8 subject areas: mathematics, Spanish, biology, physics, chemistry, social studies, philosophy, and English. However, starting in 2014, the structure of the test was changed to make its results comparable with those of other tests administered by Icfes. In this sense, it was divided into five subject areas: mathematics, science, critical reading, social studies and English. The score for each area ranges from 0 to 100, and the overall test score ranges from 0 to 500 points, calculated as a weighted average of the individual tests. Our main analysis focuses on 2015-2016 results (same structure), although we also use data prior to 2014 for some tests and robustness checks.

Owing to the importance of *Saber 11*, Icfes used to offer a familiarization test called *Pre Saber* to students who wanted to prepare for it. It costed approximately $30 USD, which corresponds to about 9% of the minimum monthly wage at the time,[24] something that low-income students cannot easily afford (Bernal & Penney, 2019). Today, the Icfes' website contains free resources that anyone can access, including mock questions for each of the subjects. In addition, private companies also offer courses with mock exams and test-oriented classes, although these are mainly used by private institutions. Students in public schools, who are usually from low-income families, often miss out on these opportunities.

2.2 Higher education

The higher education system in Colombia is comprised of public and private institutions that perform admission procedures each semester. In these processes, students apply to specific programs but are not limited to a single institution or program. There are two main types of programs that are offered by institutions: short-cycle and professional.[25] *Saber 11* plays a central role in the admission processes of these institutions (Londoño-Vélez et al., 2020), with its score being required by some of them as a selection mechanism. Other institutions such as *Universidad Nacional de Colombia* or *Universidad de Antioquia* (UdeA) use their own standardized tests as a selection mechanism.

Additionally, students in Colombia must take a standardized test to be able to graduate from a higher education institution. Its functioning is similar to that of the *Saber 11* and it is also administered by Icfes. It consists of two sections: generic competencies (which includes quantitative reasoning, critical reading, English, written communication,

[23] In Colombia, schools are either calendar A or B, which refers to the beginning and end of their academic years. For example, calendar A schools start in January and end in November, while calendar B schools start in August and end in June.

[24] Back of the envelope calculation using the average December 2014 exchange rate of 2344 COP per USD.

[25] Technical and technological programs (short-cycle) have a duration of two and three years, respectively. Some institutions, like *Servicio Nacional de Aprendizaje* (SENA), even offer technical programs with a duration of one year. Professional programs span from four to five years.

and citizenship skills) and a specific section that includes questions related to the student's major of study. The test is called *Saber Pro* for professional careers and *Saber TyT* for short-cycle programs. It is only taken by students who are at the final stage of their studies and are close to graduation.

As Ferreyra (2021) points out, the costs of higher education in Colombia are particularly high compared to other countries in the region. This is mainly explained by private professional programs, which are significantly more expensive than public ones. Although both public professional and short-cycle programs are highly subsidized by the government, the cost of studying one of these programs is higher when compared to other Latin American countries, but it is still affordable for low-income students.

As a result, the country's largest public universities have highly competitive admissions processes, and only the highest-achieving students get admitted. At private institutions, however, funding is the main channel through which low-income students can enroll. And while scholarships are mostly taken by higher-achieving students, there is a large market for educational loans in which to finance their studies. Most of this market is controlled by the *Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior* (ICETEX), a large public institution in charge of granting student loans. However, most students at high-quality private institutions are high-income individuals, while low-income students tend to opt for low-quality institutions or the National Learning Service (SENA). The latter is a public establishment attached to the Ministry of Labor that offers free training in short-cycle programs that contribute to the country's economic, scientific and social development. Low-income students tend to apply massively to SENA, so the institution has a standardized entrance test to determine who gets the available places.

2.3 *SaberEs* program

In 2016, the Secretariat of Education of the Mayor's Office of Medellin implemented a program called *SaberEs*. This initiative arose from the 2016-2019 Development Plan of the city of Medellin, specifically from component 4.2.3.1, which proposed a strategy for the development and strengthening of cognitive skills (Medellin Mayor's Office, 2016). As such, *SaberEs* aims to develop skills that strengthen preparation for standardized tests such as *Saber 11*, which in turn allow students to aspire to tertiary education scholarships and prepare for their entrance exams. To achieve this, the Development Plan establishes the use of mock tests as a familiarization and diagnostic tool for students, as well as competency-based training sessions to develop skills in the analysis and solution of the types of questions used in these tests. Additionally, the program has components like teacher training, installed capacity in educational institutions, and vocational guidance for senior students.[26]

Consequently, in the first year of the program, two firms were hired by the Secretariat of Education to carry it out in the city's public institutions. These firms were the same ones that elite private institutions had contracted for several years to prepare their stu-

[26] In the vocational guidance component, all public school students take a vocational and occupational orientation test that suggests areas of study based on their measured abilities. This test was administered one month after students took the *Saber 11* exam. Students took the *Saber 11* test on July 31, while the vocational and occupational orientation test took place on August 30.

dents. The city's budget allowed them to conduct the vocational guidance test with all public school students, but the test preparation component could only be given to 68% of the schools. Thus, each company was assigned a separate set of public schools whose selection was primarily based on their previous score on the *Saber 11*,[27] so the selected institutions could not anticipate their inclusion in the program. This implies that when ranking all public schools in the city based on the previous year's scores (2015), those at the bottom would be the ones eventually being treated.

However, when checking the selection process in our data, we find that the assignment mechanism did not work perfectly in practice. Although most treated schools were concentrated at the bottom of the distribution, some of the top-ranked schools were also selected, proving that compliance with the specific rule was not perfect. In spite of this, treated schools had lower ranks and standardized scores in the pre-treatment year (2015) than the control group on average. So altogether, the assignment was consistent with the main selection criterion described in the official documents. The key part of this process is that the schools were not informed of this decision prior to the start of the program, which is necessary for our identification strategy.

With the list of schools to be treated, each company was responsible for implementing the program in their assigned schools from eighth through eleventh grade (senior year in Colombia), yet most of their resources were concentrated on seniors. Specifically, these companies implemented the program between June and July of 2016 as follows. First, they focused on training teachers in the five subject areas of the *Saber 11*. Second, they trained school principals and coordinators in pedagogical and methodological strategies; and finally, after the students received the lessons from their teachers during school hours,[28] they took a mock test. After this test, feedback sessions were held with the students and teachers. For example, in the case of the company with almost two-thirds of the population of the assigned institutions, a total of three mock tests were given to grade 11 students, each of which was subsequently reviewed in the feedback sessions. This was followed by teacher training sessions covering one of three main topics in the following order: 1) the referents of the PISA international assessment, 2) the review and substantiation of the curricular components (curricula, contents, competencies) of the evaluated subjects, and 3) critical reading as a transversal axis of knowledge.[29]

It was not long before the press and politicians anecdotally claimed that the program had a positive impact, significantly improving public school test scores in 2016.[30] Furthermore, the vocational guidance component was also acclaimed for providing students with enough information to make a more informed decision about their post-secondary education.[31]

[27]The first company, *Los Tres Editores*, was assigned a total of 98 institutions, while the second company, *Avancemos*, was assigned 52. It was only up to 2019 that all public institutions were covered.

[28]This is key to our identification. If the program was an extracurricular activity that happened outside school hours, effects could be confounded by the fact that students have more schooling hours.

[29]These training sessions had perfect compliance rates.

[30]`https://bit.ly/3ovyYSW`

[31]`https://bit.ly/3PGQxv4`

3 Data

We use administrative records from eight main sources.[32] First, *SaberEs* treatment data comes from the Secretary of Education of the Mayor's Office of Medellin. Based on public contracts, we were able to identify 150 public schools out of 220 that received the program in 2016. In addition, we collected information on the schools that received the program in 2017.[33]

Our second database, and baseline data, come from the *Instituto Colombiano para la Evaluación de la Educación* (Icfes), the institution responsible for measuring the quality of education through standardized testing in Colombia. In addition to the *Saber 11* scores for all students in the country, these data contain detailed sociodemographic information (e.g., socioeconomic stratum, parental education) for all *Saber 11* test-takers. By merging the administrative records of the program with the Icfes data,[34] we can identify students in treated and control schools and are able to access a large set of baseline covariates. Since the program was constrained to the public schools in the city of Medellin, we restrict our main sample to this population.

Our main sample is merged with six additional datasets. First, we use the Ministry of Education's *Sistema Nacional de Información de la Educación Superior* (SNIES), which follows students in their postsecondary education. We use SNIES data from 2016 to 2019,[35] which provide student-by-year-level information on characteristics such as enrollment status on postsecondary education and major choice. We focus on professional and short-cycle students who are enrolled in a higher education institution for any given semester in the aforementioned time frame. It is important to acknowledge that in this case we cannot include more pre-treatment periods and perform the same dynamic analysis that we do with the *Saber 11* scores.[36]

Second, we use information from the *Saber TyT* dataset from Icfes. Here, we use data for all *Saber TyT* test-takers in the spring and fall semesters of 2016 through 2019. Given that *Saber TyT* is a mandatory exam for all short-cycle students in Colombia, this source provides us with a proxy for graduation from short-cycle programs.

Third, we use data from ICETEX, the institution that manages scholarships and financial aid in Colombia on behalf of public and private organizations. On one side, we have administrative data on national financial aid at the individual level for 2018 and 2019. This dataset contains all student's financial aid loans financed with public funds that offer below-market interest rates. In order to access these credits, students need

[32] Details of our approach to data construction are included in Appendix A.

[33] Our main analysis focuses on public schools who were the potential beneficiaries of the program. So, we removed the observations of private schools to ensure valid comparisons. We only used private schools to calculate the average rank of their students and have an additional measure of the gap with respect to the treated public schools. In addition, private schools are problematic from the technical point of view, since most of them offer their students preparatory courses for *Saber 11*.

[34] To merge all datasets, we use national identification numbers, names, and date of birth (see Appendix A for details.

[35] We excluded 2020 data due to the COVID-19 pandemic, since there was atypical college enrollment behavior.

[36] Prior to 2016, the higher education data were stored in another system called SPADIES. Since the two datasets are fundamentally different from each other in the way the information was collected, in particular with respect to short-cycle programs, we decided to work exclusively with SNIES to ensure adequate estimates.

to meet certain threshold scores on the *Saber 11*. However, these funding opportunities are offered at a national level and may not be the primary option for all students. For example, there are local government agencies that largely cover scholarship supply at the municipal or state level. In the case of Medellin, Sapiencia is the main local institution that manages scholarships and financial aid. We use Sapiencia's data from 2016 to 2019 to identify beneficiaries. With the two datasets mentioned above we constructed a variable indicating whether the student received any type of financial aid (either from ICETEX or Sapiencia).[37]

Fourth, we track access to *Ser Pilo Paga* (SPP), a nationwide scholarship that annually financed the entire undergraduate education in any private or public institution in Colombia for 10,000 students from low-income households who scored above the 90th percentile on the *Saber 11*. We focus on beneficiaries between 2015 and 2016, as individuals were only eligible in the same year they took the *Saber 11* for the first time. Therefore, we created a dummy variable indicating whether a student received the SPP scholarship to observe the effects on access to a merit-based scholarship.

Finally, we use data from *Olimpiadas del Conocimiento*, a competition that the Secretary of Education carries out each year between all private and public school students in Medellin.[38] We use this dataset to test one of the hypothesized mechanisms. Here, we use individual-level data on test scores from the 2015 and 2016 versions of the contest to check whether *SaberEs* had an impact on them. As such, we constructed a unique dataset containing granular information on students' socioeconomic characteristics, performance on high school exit exams, access to all types of higher education, performance on college exit exams, and nearly-global access to financial aid and scholarships at the individual level.

3.1 *Saber 11* scores

To make scores comparable, we follow Laajaj, Moya, and Sánchez (2022) and use the student's rank within their cohort as the variable of interest and rescale it from 0 to 100 (worst to best, respectively). This ensures comparability of scores between years even after the difficulty or structure of the test changes. However, the entire analysis is also performed with standardized test scores as a robustness check.

As mentioned in the section 2, *Saber 11* had a structural change in 2014, making test scores uncomparable across all periods in our sample. For this reason, we consider 2015 and 2016 to calculate the main effects. Most of the analysis is performed using what is commonly referred to as a 2x2 difference-in-differences specification in order to obtain the most accurate estimates of the program's impact under the least strict assumptions. Nonetheless, we extend our sample to the period 2010-2017 in order to provide formal and informal tests of the parallel trends assumption and to provide further robustness checks of the effect of *SaberEs* on *Saber 11* scores. Especially since extending our timeline from 2010 to 2017 would introduce dynamic effects, which combined with the staggered adoption nature of the treatment would impose additional assumptions and require other methods to ensure the validity of the estimation. We discuss such methods in the next

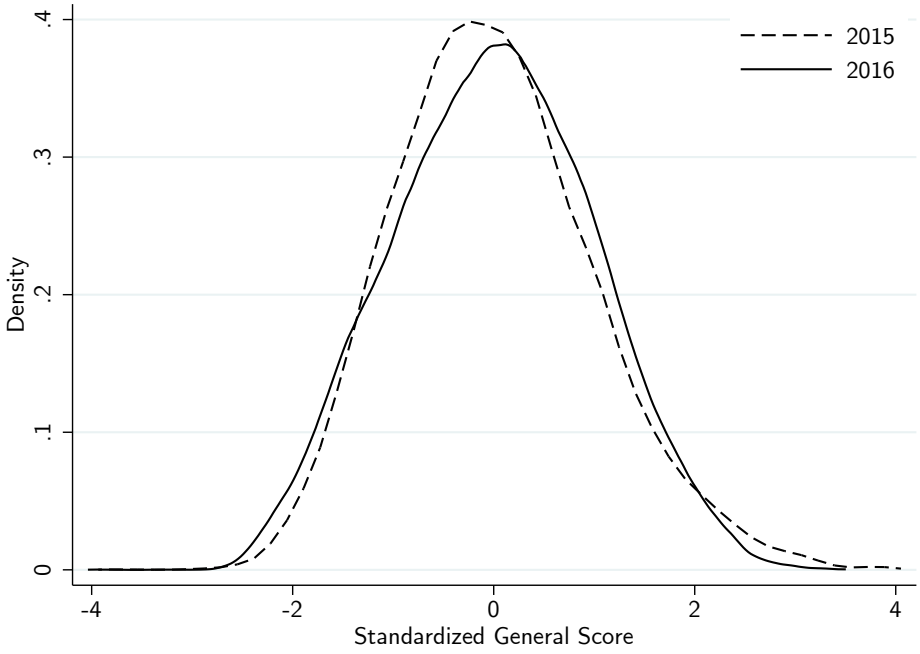
[37] We focus on this variable for simplicity, but the results are the same when separating our estimation into ICETEX and Sapiencia access individually.

[38] The details of the competition are found in section 6.

section.

As a purely descriptive exercise, Figure 1 shows the distribution of general standardized scores for 2015 and 2016. There is an increase in test scores in the treatment year, although it appears to be concentrated at the top-half of the distribution. This descriptive analysis points to the presence of heterogeneous treatment effects across the score's distribution, which is more formally analyzed in section 5.

Figure 1: Standardized General Score Density 2015-2016.



The summary statistics for the 2x2 sample (2015-2016) are shown in Table 1. Panel A shows unstandardized test scores. Panel B shows general access to higher education and access by program. 60% of students who graduated in 2015 and 2016 accessed any type of higher education within three years of graduation. Panel C shows the treatment variables; in this case, 66% of students were treated by at least one of the companies, with 46% being treated by *Los Tres Editores* and 20% by *Avancemos*. Finally, Panel D contains all socioeconomic covariates obtained from the *Saber 11* dataset. Around 57% percent of the students were female and only 10% had at least one parent with some tertiary education. Consistent with the sample involving public schools, only 4% of the students' households were of high stratum[39] and 7% of them had a high income (defined as household income above three monthly minimum wages).

[39]In Colombia, households are divided into one of six strata depending on their area of residence. This reflects their socioeconomic status, where the sixth stratum is the highest.

Table 1: Summary Statistics 2015-2016

	Mean	SD	Min	Max
Panel A: Test Scores				
General	258.70	42.14	13	450
Reading	52.85	8.93	0	100
Math	51.13	10.52	0	100
Science	51.48	9.11	0	100
Social studies	51.52	10.06	0	93
English	51.66	10.37	0	100
Panel B: Higher Education and Financial Aid				
Access to higher education	0.60	0.49	0	1
Access to short-cycle	0.31	0.46	0	1
Access to university	0.33	0.47	0	1
Access to STEM	0.26	0.44	0	1
Access to professional STEM	0.16	0.37	0	1
Access to short-cycle STEM	0.13	0.34	0	1
Received financial aid	0.05	0.22	0	1
Received *Ser Pilo Paga*	0.03	0.16	0	1
Panel C: Treatment				
Treated	0.66	0.47	0	1
Treated Tres Editores	0.46	0.50	0	1
Treated Avancemos	0.20	0.40	0	1
Panel D: Covariates				
Female	0.57	0.50	0	1
TV	0.80	0.40	0	1
Oven	0.60	0.49	0	1
Landline	0.85	0.36	0	1
Microwave	0.50	0.50	0	1
PC	0.78	0.42	0	1
Car	0.16	0.37	0	1
Internet	0.77	0.42	0	1
Washing machine	0.81	0.39	0	1
DVD	0.61	0.49	0	1
NSE 1	0.03	0.18	0	1
NSE 2	0.33	0.47	0	1
NSE 3	0.62	0.48	0	1
NSE 4	0.02	0.13	0	1
Employed	0.06	0.23	0	1
Parent's education	0.10	0.30	0	1
High income	0.07	0.26	0	1
High stratum	0.04	0.20	0	1
Household floor	0.42	0.49	0	1
$>$ 6 People in household	0.20	0.40	0	1
$>$ 3 Rooms in household	0.61	0.49	0	1

Notes: NSE is the socioeconomic level of the student (NSE), given that Icfes classifies students into four levels (the fourth one being the highest) according to their parent's education and occupation, as well as the family's income. Parent's education takes the value of 1 if one of the parents has some tertiary education (complete or incomplete). High income takes the value of 1 for individuals whose household income is above three monthly minimum wages. High stratum takes the value of 1 for households above the third stratum. Household floor equals 1 if the house's floor is made of cement, gravel, bricks, soil or sand. The rest of them are self-explanatory.

4 Empirical Strategy

To identify the causal effect of *SaberEs* on students' test scores and other educational outcomes, we exploit the timing of the program and the lack of universal coverage using difference-in-differences (DiD) and a two-way fixed effects estimation. First, we focus on the simple 2x2 case where there is no staggered treatment adoption. In this case, we estimate a simple DiD regression as follows:

$$Y_{ict} = \alpha_0 + \alpha_1 \text{Treated}_c + \alpha_2 \text{Post}_t + \beta \text{Treated*Post}_{ct} + X'_{ict}\delta + \varepsilon_{ict} \quad (1)$$

where Y_{ict} is the outcome of student i from school c at time t, $Treated_c$ indicates whether school c is treated, $Post_t$ takes a value of 1 if the student's test application year is 2016, and $Treated*Post_{ct}$ is their interaction. Finally, X'_{ict} is a vector of controls that contains socioeconomic covariates taken from the registration form of the *Saber 11* test, and ε_{ict} is the error term. In all our exercises (unless otherwise stated), we cluster standard errors at the school level.

The coefficient of interest is β, which under the assumption of parallel trends captures the effect of the *SaberEs* program. In our case, this implies that the evolution of test scores between the treatment and control groups over time is parallel and would continue as such if the treatment never took place (counterfactual scenario). We estimate Equation 1 with and without controls to test the robustness of the results after their inclusion, understanding that in that case we are implicitly assuming that parallel trends hold conditional on the covariates and that there are no heterogeneous treatment effects.

Additionally, we estimate the following two-way fixed effects regression (TWFE):

$$Y_{ict} = \alpha_0 + \beta \text{Treated*Post}_{ct} + \psi_c + \gamma_t + \epsilon_{ict} \quad (2)$$

where ψ_c and γ_t are the school and year fixed effects, respectively, and ϵ_{ict} is the error term. Notice that TWFE accounts for unobserved characteristics of schools as well as for any constant effect over time.

The inclusion of additional covariates imposes additional assumptions that relate to the controls' specific trends and to the homogeneity of treatment effects, so it could more likely lead to biased estimates of the underlying effects (Sant'Anna & Zhao, 2020). In order to take advantage of the large set of available covariates without imposing additional assumptions, we implement the Hájek (1971) type inverse probability weighting (IPW) estimator that normalizes weights to sum up to one -which is more stable-, and the outcome regression (OR) (Heckman et al., 1997). Moreover, we use Sant'Anna and Zhao (2020) doubly robust difference-in-differences regression to obtain a more efficient and reliable estimator. This method combines OR and IPW to come up with an estimation that is robust as long as at least one of the two models is correctly specified, therefore allowing for more flexibility. Specifically, we focus on the improved estimator for repeated cross-sections based on the structure of our data (Sant'Anna & Zhao, 2020).

In addition to the 2x2 case described above, we also analyze a non-staggered dynamic specification in which we use observations from 2010 to 2016 to estimate the causal effect of the program on the students' scores. Specifically, we use TWFE to estimate all coefficients before and after the program and provide informal evidence of the parallel trends assumption. Furthermore, we include units treated in 2017 in our sample period. Since

there are now multiple time periods and two years of treatment (i.e., there is staggered adoption), a simple TWFE regression would potentially be biased due to the presence of heterogeneous effects (Borusyak & Jaravel, 2017; De Chaisemartin & d'Haultfoeuille, 2020). This happens because the estimator is a weighted average of all 2x2 comparisons and therefore includes "forbidden comparisons" that could have negative weights and change the sign of the estimate (Goodman-Bacon, 2021).

To overcome this challenge, we use the Callaway and Sant'Anna (2021) estimator, which calculates all group-time specific effects. The procedure also allows for aggregations to be made in an "event study" form and in a simple one that reports a single coefficient. In particular, we calculate both aggregations and also check the robustness of our results using an alternative specification proposed by Borusyak et al. (2021), even though it relies on a stronger assumption about parallel trends and could lead to a larger bias if it does not completely hold (Roth et al., 2022). With this in mind, we focus on the simple aggregation as recommended by Callaway and Sant'Anna (2021), but we also present the event study aggregation.

4.1 Identification strategy

The difference-in-differences (DiD) method has become a widely used empirical strategy for identifying the causal effects of policies and interventions in economics. This is mostly due to its intuitive interpretation and its ability to uncover causality even in non-experimental settings. However, like all empirical methods, the validity of DiD estimates relies on some key assumptions being met. In this subsection, we discuss the identification strategy of our study, focusing on the key assumptions of the DiD method and how we have attempted to address them.

The first two assumptions of a canonical DiD model are the no anticipation assumption, and the stable unit treatment value assumption (SUTVA). The former requires that the treatment group would not have anticipated the policy intervention and adjusted their behavior accordingly before it happened. If there was anticipation, it would bias the DiD estimates. On the other hand, SUTVA requires that the policy intervention solely affects the treatment group and does not spill over to the control group or vice versa. This means that there should be no interference between the units, and each unit's outcome should depend only on its own treatment status and not on the treatment status of other units.

In our context, schools were unaware that their inclusion in the program was based on their prior performance on the *Saber 11* exam, which was used as the main selection criterion. As a result, schools could not have anticipated their treatment status and were unable to strategically adjust their behavior in response to the program. In addition, the vocational guidance component does not affect our identification strategy since the vocational and occupational orientation test was administered one month after students took the *Saber 11* exam. In particular, *Saber 11* was administered on July 31, while the vocational test was taken on August 30.

Regarding SUTVA, the *SaberEs* program was implemented at the school level, meaning that all students within a given school were treated simultaneously. Additionally, the fact that test preparation occurred during regular school hours further reduces the likelihood of spillover effects, making it unlikely that the program's impact spilled over

to the control group or to other schools.

Finally, another key assumption is parallel trends, which requires that in the absence of *SaberEs*, scores would have followed the same trajectory (trend) in the treatment and control groups. This assumption is essential for ensuring that any observed differences between the treatment and control schools after the intervention are attributable to *SaberEs* and not due to pre-existing differences in the groups. In practice, the most common assessment is to look at the significance of the pre-trends in an event study estimation, what is known as pre-trends testing. If these coefficients are not statistically significant, it favors the validity of the assumption.

In our 2x2 specification, it is not possible to test for parallel trends since there is only one pre-treatment period. However, to provide additional evidence for this assumption, we extend the sample to perform dynamic estimations as a robustness check. Furthermore, we conduct an event study estimation that shows non-significant pre-trends, which is consistent with the Callaway and Sant'Anna (2021) estimates. To further assess the validity of the parallel trends assumption, we also calculate the possible bias from pre-testing following the approach of Roth (2022) and perform a sensitivity analysis to test the robustness of the results to linear and non-linear violations of parallel trends, as proposed by Rambachan and Roth (2023).

5 Results

This section presents the main results organized as follows. In subsection 5.1, we show the effects of *SaberEs* on the students' scores using the 2x2 case. Then, subsection 5.2 contains the dynamic specification, where we extend our sample from 2010 to 2017 and provide additional evidence on the effects of the program on scores. In addition, subsection 5.3 shows the heterogeneous effects on the outcome's distribution using unconditional quantile regression (Firpo et al., 2009). Finally, in subsection 5.4 we display the effects on higher education by type of program.

5.1 Effects on students' score

Table 2 reports the effects of *SaberEs* on students' test scores using different approaches. Columns 1-6 show the results for the student's general rank on the *Saber 11* and Columns 7-12 show the results for the standardized scores. All estimates are calculated at the end of high school (11th grade) and the coefficients in the table are directly interpretable as increases in the student's rank in the test, or as standard deviations from the mean in the case of standardized test scores.

In Table 2, Column 1, we report the standard DiD estimator without controls using Equation 1. We find that *SaberEs* increases the student's general rank by 2.97. When including the set of controls (Column 2), the effect adjusts slightly downward, but remains at a significant 2.6 ranks. The two-way fixed effects specification (Column 3) shows a smaller point estimate, yet the effect follows the same pattern. In fact, coefficients across the six specifications are not statistically different from each other. When looking at more robust specifications, such as outcome regression and inverse probability weighting with stabilized weights (Columns 4-5), the effect remains statistically significant and similar

to the previous estimates. Finally, column (6) shows the result for the doubly robust DiD estimator (Sant'Anna & Zhao, 2020), our preferred option, which highlights an increase in the general rank of 2.2 ranks.

The results on standardized test scores follow the same pattern, showing that *SaberEs* increased the average student's test scores by around 0.07-0.1 standard deviations (0.074 in the most robust specification). As with the student's rank, there is no statistically significant difference in the coefficients across specifications.

In each case, to make the result more easily interpretable, the third row reports the coefficient in terms of the gap reduction relative to the control group. This comes from a back of the envelope calculation in which we divide the estimated coefficient by the difference in the average rank of untreated and treated students in 2015.[40] We find that the program generated a 22.9% reduction in the rank's gap between treated and untreated students. Gap reductions with other comparison groups can be found in Table B.1 in the appendix.

Table 2: Main results

	Student's rank						Standardized test scores					
	(1) DiD	(2) DiD	(3) TWFE	(4) OR	(5) IPW	(6) DR	(7) DiD	(8) DiD	(9) TWFE	(10) OR	(11) IPW	(12) DR
SaberEs effect (β)	2.965***	2.559***	1.862**	2.222**	2.693***	2.222**	0.104***	0.089***	0.066**	0.073**	0.092**	0.074**
	(0.976)	(0.886)	(0.827)	(0.917)	(1.032)	(0.915)	(0.034)	(0.030)	(0.029)	(0.032)	(0.036)	(0.032)
Gap reduction	30.6%	26.4%	19.2%	22.9%	27.8%	22.9%	31.1%	26.6%	19.7%	22.0%	27.4%	22.2%
Observations	35,495	35,484	35,495	35,484	35,484	35,484	35,495	35,484	35,495	35,484	35,484	35,484
Controls	NO	YES	NO	YES	YES	YES	NO	YES	NO	YES	YES	YES

Notes: Standard errors clustered at the school level. The different specifications are, in their respective order: Difference-in-Differences (DiD) without controls, Difference-in-Differences (DiD) with controls, Two-Way Fixed Effects (TWFE) without controls, Outcome Regression (OR), Inverse Probability Weighting (IPW) with stabilized weights, and Improved Doubly Robust Difference-in-Differences (DR) for repeated cross-section. The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

As seen in Figure B.1 and Figure B.2 in the appendix, the positive effect is being driven primarily by an increase in math, science, and social studies, regardless of the outcome variable (rank or standardized test scores). The critical reading test also appears to be driving the results, but the effect is not robust if standardized scores are used as the outcome variable, then we cannot fully conclude its relevance. Regarding the English test, the effect is not significant, which is to be expected given that the program was not designed to teach a second language and could hardly have done so in such a short time.

These results are especially important when compared to what has been found in related studies. For example, Laajaj et al. (2022) finds that *Ser Pilo Paga* created a motivational effect that caused individuals to accumulate human capital and, therefore, increased the student's rank by 1.57, which was the case for students at the top of the distribution. In our case, we found similar effects in all specifications. On the other hand, Gómez et al. (2020) finds that test preparation activities increased the general score by around 0.07 standard deviations, although they were more useful for students in private schools and when performed during after-school hours. Likewise, Bernal and Penney (2019) finds that the SPP program led to a 0.09 standard deviation increase in test scores for eligible students, and the effects were also concentrated at the top of the distribution. As reported in Table 2, Column 12, we find that *SaberEs* increased

[40]The calculation is performed as $\beta \,/\, E[Y_0 - Y_1 \mid t = 2015]$.

test scores by around 0.074 standard deviations, which is a similar effect to what these studies found. Yet in this case it is not solely driven by the very top-performers as shown in subsection 5.3.

5.2 Dynamic specifications

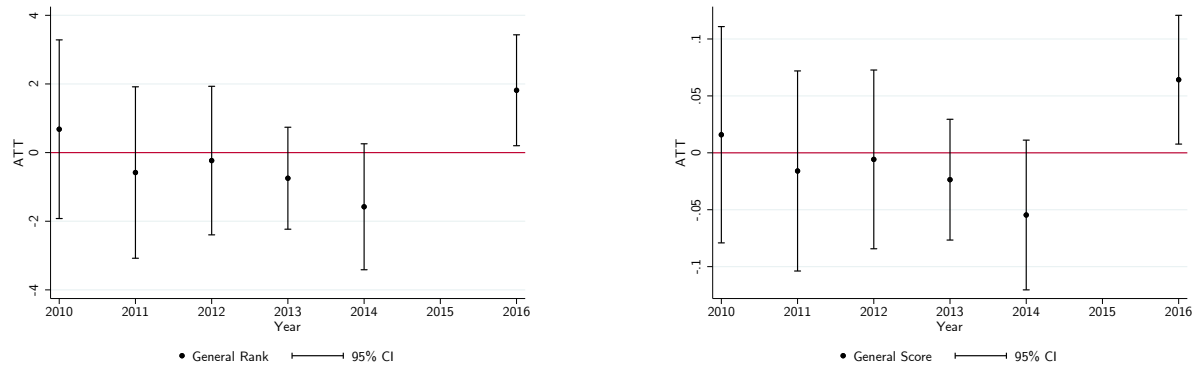
The 2x2 specification shown in the previous section relies on the assumption that trends are parallel to each other. In this subsection, we not only check for the presence of possible pre-trends but also test if our results hold after increasing the sample and introducing staggered treatment adoption.

First, we extend our sample from 2010 to 2016, and estimate a dynamic DiD using two-way fixed effects as:

$$Y_{ict} = \alpha_0 + \sum_{h=2010}^{2016} \beta_h[t * \text{Treated}_{ch} = h] + \psi_c + \gamma_t + u_{ict} \quad \forall \quad h \neq 2015 \tag{3}$$

where Y_{ict} is the outcome of student i from school c at time t, and $t*Treated_{ct}$ are the interactions of year and treatment status for each of the leads or lags (h). ψ_c and γ_t are the school and year fixed effects respectively, and u_{ict} is the error term. Standard errors are clustered at the school level. Our results are consistent with the 2x2 case. Figure 2 shows a positive effect of *SaberEs* on scores in the year of treatment. In addition, it shows no discernible impact before the treatment took place for either the general rank or the standardized test scores. This informally implies that there were no significant pre-trends in the design. In general, the program had a positive effect on the students' rank that corresponds to a gap reduction of over 24% with respect to the control group (see Table B.2). Results are robust to using ranks or standardized scores.

Figure 2: Event study estimates with non-staggered treatment adoption



(a) Student's rank

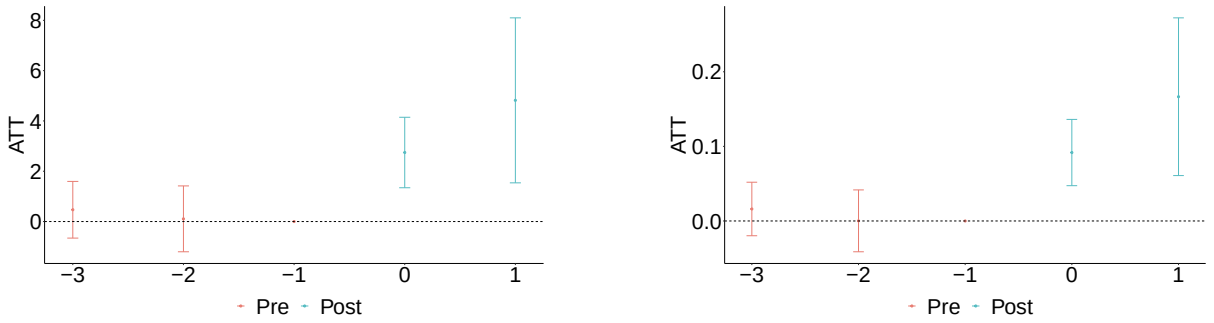
(b) Standardized test scores

Note: Standard errors clustered at the school level. Estimates come from an event study estimation where the first lead (2015) is the omitted category.

Moreover, we extend the analysis to 2017, when a new group of schools entered the program. Since there are two years of treatment with multiple time periods, we estimate

a dynamic DiD with staggered adoption. We start by presenting the results as an event study. In this case, in order to avoid the issue of compositional changes, we present the results using balanced groups around the event time, as suggested by Callaway and Sant'Anna (2021). The results are consistent with our previous findings. These are presented in Figure 3 for the student's rank and the standardized scores. There is a positive and statistically significant effect of the program even after allowing for staggered treatment adoption by including 2017. On the other hand, the coefficients for the pre-treatment periods are not statistically different from zero, which at first glance might indicate the absence of pre-trends. Overall, the coefficients point to a similar conclusion -highlighting the robustness of our results- and are slightly larger than those obtained in the 2x2 case. This is consistent with the fact that the program targets eighth to eleventh grade students, so the effects are capturing the fact that tenth grade students received this training twice and experienced greater rank improvements when they took the test.[41]

Figure 3: Event study results with staggered treatment adoption



(a) Student's rank

(b) Standardized test scores

Notes: Estimation based on Callaway and Sant'Anna (2021). Results are presented with balanced groups around the event time to avoid the issue of compositional effects, as suggested by the authors. Controls include gender, household goods and services (computer, car, internet and washing machine), parents education, and stratum.

Table 3 aggregates the coefficients to a single effect following the approaches proposed by Callaway and Sant'Anna (2021) and Borusyak et al. (2021). The first two columns for each outcome show the results without controls, while the last two report the results with the inclusion of controls. Overall, the coefficients point to a similar conclusion. In terms of the gap reduction relative to the control group, the program reduced the gap between 28% and 38%, as it had a positive and significant effect of around 3.6 ranks in the most robust specification. This is also the case when looking at the results using standardized test scores as the variable of interest, as the gap reduction relative to the control group is between 30% and 40%.

[41] *Caveat:* In this case, we are not only dealing with staggered treatment adoption, but there are different intensities of treatment between the two years. Our estimate could possibly be a lower bound of the results.

Table 3: Dynamic Results

	Student's rank				Standardized test scores			
	(1) C&S	(2) BJS	(3) C&S	(4) BJS	(5) C&S	(6) BJS	(7) C&S	(8) BJS
SaberEs effect (β)	3.715***	2.711***	3.598***	2.688***	0.131***	0.099***	0.123***	0.094***
	(0.785)	(0.497)	(0.820)	(0.462)	(0.028)	(0.016)	(0.029)	(0.015)
Gap reduction	38.3%	28.0%	37.1%	27.7%	39.2%	29.6%	36.8%	28.1%
Observations	147,656	147,554	147,656	70,859	147,656	147,554	147,656	70,859
Controls	NO	NO	YES	YES	NO	NO	YES	YES

Notes: Standard errors clustered at the school level. C&S relates to the "simple" aggregation from Callaway and Sant'Anna (2021). BJS relates to the estimator proposed by Borusyak et al. (2021). Controls include gender, household goods and services (computer, car, internet and washing machine), parents education, and stratum. *p<.05; **p<.01; ***p<.001

Although the estimates from the event study above provide an informal test for the parallel trends assumption, we further analyze the pre-trends by conducting a power and sensitivity analysis. First, we test whether the pre-treatment trends are parallel with a formal test as suggested by Roth (2022), which can be seen in Table B.3. Here, using the precision of the estimates, we compute the pre-trend that has 50% power of being detected (hypothesized trend) and an adjusted pre-trend that considers the bias generated from an analysis being done conditional on passing a pre-test under the hypothesized trend. Based on the likelihood ratio we can conclude that estimating coefficients similar to the ones we observe is more likely under parallel trends than under the hypothesized linear trend.

Finally, we conduct a sensitivity analysis based on Rambachan and Roth (2023), where we estimate a 95% confidence set for the general rank to consider its robustness to some degree (M) of deviation from the parallel trends assumption. Specifically, we check for linear (M=0) and non-linear (M>0) deviations. Figure B.3 reports the confidence set that results from this estimation. We find that our results are significant when allowing for a linear extrapolation of the pre-existing trend. Additionally, when allowing for non-linear deviations we find that the increase in the students' general rank is robust. This is explained by the magnitude of the breakdown value of M, which in this case is more than four times larger than the size of the pre-trend that has 50% power of being detected as shown in the power analysis previously described.[42]

The fact that our results are robust to large non-linear deviations from parallel-trends as well as to the power analysis, further proves the reliability of our estimates. When conducting both of these analyses using the standardized test scores as the variable of interest (see Table B.3 and Figure B.4), we can conclude the same.

5.3 Heterogeneous effects on the score's distribution

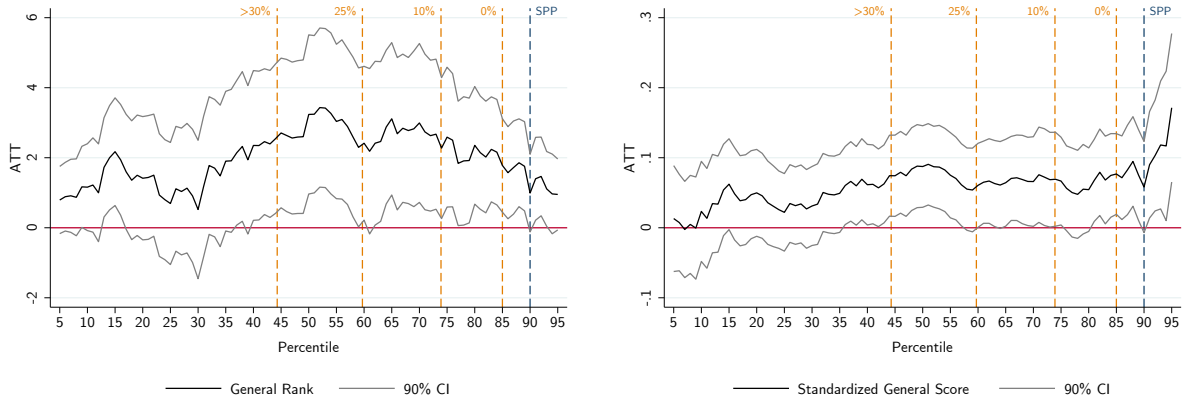
It is also important to look at the treatment effects of *SaberEs* throughout the outcome's distribution. Given that OLS regressions yield the effects on the unconditional mean, we use recentered influence function (RIF) regressions (Firpo et al., 2009) to examine what happens at the unconditional quantiles. This will allow us to determine whether

[42]The 50% pre-trend corresponds to the dashed blue line in the figures.

the effects are concentrated among high- or low-performing students. Figure 4 shows the results of the RIF regressions on the percentiles of the test's rank and standardized scores distributions for the 2x2 scenarios. The figure also shows some of the cut-off points for financial aid opportunities. Specifically, four types of ICETEX credits and *Ser Pilo Paga* (SPP).

Effects are statistically significant not only for students around the 50th percentile, but for the top-performers. Thus, some of the effects are focused on students above the financial aid thresholds, implying the possible presence of effects on access to these opportunities (we check this in subsection 5.4).

Figure 4: Effects by students' rank percentiles (2015-2016)



(a) Student's rank

(b) Standardized test scores

Notes: Estimates come from unconditional quantile regressions (Firpo et al., 2009), with standard errors clustered at the school level. The yellow lines represent ICETEX credits. The percentages on top reflect the share of the total credit amount that the students are required to pay while they are studying. The lower the share, the higher the required percentile to access that credit. The blue line at the 90th percentile represents the cut-off for *Ser Pilo Paga*.

5.4 Effects on access to higher education

Now that we have established that *SaberEs* had a positive impact on the students' rank, it is important to see whether this effect translates into increased access to higher education. In this sense, we estimate a doubly robust difference-in-differences regression as in Sant'Anna and Zhao (2020).[43] Table 4 shows the results for access to higher education in general, and divided by type of program: professional (five-year programs) and short-cycle (two- or three-year programs). Column 1 shows that the program had a positive effect on access to any form of higher education in the first year after the program of 3.3 percentage points, representing an effect of 6.5% relative to the control group. The point estimate is similar for the first two years after the program, although for the last year it is slightly lower. These effects are mainly due to the greater access to short-cycle programs. The effect on short-cycle programs is significant up to two years after graduation, which is

[43]We are not able to do the dynamic estimations for higher education since we only have data from 2016 onward. Recall that SNIES started in that year, and prior to that, data were collected in SPADIES.

consistent with their two-year duration. In the case of professional degrees, the program has no perceivable impact, although the precision increases significantly for the last year.

Table 4: Effects on access to a higher education program

	Higher Education			Short-cycle			Professional		
	(1) 1 year	(2) 2 years	(3) 3 years	(4) 1 year	(5) 2 years	(6) 3 years	(7) 1 year	(8) 2 years	(9) 3 years
SaberEs effect (β)	0.033**	0.039***	0.024**	0.026**	0.023**	0.010	0.006	0.015	0.014
	(0.015)	(0.015)	(0.012)	(0.013)	(0.011)	(0.011)	(0.011)	(0.012)	(0.010)
Observations	35,484	35,484	35,484	35,484	35,484	35,484	35,484	35,484	35,484
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES
Mean Control 2015	0.511	0.559	0.587	0.219	0.230	0.224	0.292	0.329	0.361

Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). The columns indicate access to each outcome 1, 2 and 3 years after students graduate from high school. The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

These results are impressive for such a low-cost program and are in line with the objective behind it. Although we cannot test this hypothesis, we believe that the vocational orientation component appears to have been successful in providing students with sufficient information to enroll in higher education. This is probably due to the vocational test that all public school students undertook, which highlighted students' abilities and gave them a list of options of career majors in which they were most likely to succeed. This, combined with their improved scores on the *Saber 11*, seems to have upgraded their self-esteem and academic expectations.

In addition to the latter, we test if the effects are also present on access to STEM majors,[44] which have been found to produce higher returns (Kinsler & Pavan, 2015; Webber, 2016). Here, we also calculate the results for professional and short-cycle STEM degrees. For this purpose, we estimate a doubly robust regression of the effects of the program on whether the student ever accessed a STEM program (in general), and whether the enrollment was to a professional or short-cycle program. Table 5 shows that *SaberEs* increased access to STEM programs, but this was primarily driven by its positive effect on short-cycle STEM majors.

[44]Science, Technology, Engineering and Mathematics.

Table 5: Effects on access to STEM programs by degree type

	(1) STEM	(2) Professional STEM	(3) Short-cycle STEM
SaberEs effect (β)	0.021*	0.011	0.019**
	(0.011)	(0.009)	(0.008)
Observations	35,484	35,484	35,484
Controls	YES	YES	YES
Mean Control 2015	0.287	0.173	0.152

Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

Now that we know effects are mostly concentrated on short-cycle programs, we would like to follow up on these students and see if they eventually graduated from their higher education institutions. We use data on *Saber TyT* from 2016 to 2019 and estimate a doubly robust regression on the probability of taking *Saber TyT*, our proxy for graduation from short-cycle programs. Table 6 shows that *SaberEs* increased graduation (as measured by participation in the *Saber TyT*) from short-cycle programs in general by 2.3 percentage points from a baseline level of 15.2%. This was also the case for short-cycle STEM programs with an effect of one percentage point from a baseline level of 6.2%.

Table 6: Effects on taking the Saber TyT by type of program

	(1) All short-cycle programs	(2) Short-cycle STEM programs
SaberEs effect (β)	0.023**	0.010*
	(0.010)	(0.006)
Observations	35,484	35,484
Controls	YES	YES
Mean Control 2015	0.152	0.062

Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

6 Potential mechanisms

In this section we discuss several potential mechanisms through which the program ends up affecting access to higher education. In the previous section we showed that the program effectively increased the *Saber 11* scores and access to tertiary education. One possible explanation is that the program fulfilled the promise of developing skills in the

analysis and solution of standardized tests. If that is the case, one potential mechanism is the accumulation of task-specific human capital (Gibbons & Waldman, 2004), which is especially helpful when students take high-stakes standardized tests. This type of human capital has been found to be an important source of improvement for individuals (Gathmann & Schönberg, 2010; Ost, 2014).

To test this idea, we take advantage of a feature of the educational market in Colombia. The largest public higher institutions, located in Medellin, such as *Universidad de Antioquia* (UdeA), *Universidad Nacional*, and SENA, require applicants to take their own standardized test as part of their admission processes.[45] We look at the effects on access to these institutions versus institutions that do not require their own standardized exams.

Although we do not find effects on access to professional programs, we find that the access to short-cycle programs in institutions with these tests increased significantly. On the contrary, we do not find any significant effects on institutions that do not have admission exams. Table 7 shows the effects on short-cycle programs once we split them between those who require an admission exam and those who do not. Access to short-cycle programs in institutions with admission exams was increased during the first two years post graduation, but there are no perceivable effects on institutions that do not require these types of exams. When looking at access to professional programs in Table 8, we see that the program had no impact in institutions without admission exams, as expected. Yet there are also null effects when looking at their counterpart.

Table 7: Effects on access to short-cycle programs in institutions with *Saber 11*-like admission exams

	Admission exam			No admission exam		
	(1)	(2)	(3)	(4)	(5)	(6)
	1 year	2 years	3 years	1 year	2 years	3 years
SaberEs effect (β)	0.026**	0.024**	0.010	0.000	-0.001	0.000
	(0.012)	(0.010)	(0.010)	(0.007)	(0.006)	(0.008)
Observations	35,484	35,484	35,484	35,484	35,484	35,484
Controls	YES	YES	YES	YES	YES	YES
Mean Control 2015	0.132	0.141	0.126	0.0865	0.0896	0.0982

Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). Institutions with *Saber 11*-like admission exams that offer short-cycle programs are SENA and UdeA. The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

[45] SENA and UdeA offer short-cycle programs (the former having the largest supply), while only UdeA and *Nacional* offer professional ones.

Table 8: Effects on access to professional programs in institutions with *Saber 11*-like admission exams

	Admission exam			No admission exam		
	(1) 1 year	(2) 2 years	(3) 3 years	(4) 1 year	(5) 2 years	(6) 3 years
SaberEs effect (β)	0.006	0.009	0.007	0.000	-0.014	0.007
	(0.005)	(0.005)	(0.005)	(0.009)	(0.012)	(0.009)
Observations	35,484	35,484	35,484	35,484	35,484	35,484
Controls	YES	YES	YES	YES	YES	YES
Mean Control 2015	0.0621	0.0745	0.0788	0.230	0.670	0.282

Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). Institutions with *Saber 11*-like admission exams that offer short-cycle programs are *Nacional* and UdeA. The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

These effects can be the result of task-specific human capital but also of motivational effects. Under motivational effects, the students' aspirations change and they are incentivized to accumulate human capital by exerting higher effort (Laajaj et al., 2022). This hypothesis can be checked through a simple test in which we look at the effects of the program on the students' performance on a low-stakes exam. The logic of it is quite simple. If students are being motivated to learn and accumulate human capital, you expect to see an improvement in their overall scores regardless of the relative importance of the exam.

To test this idea, we use a low-stakes exam that is unique to the city of Medellin. Every year the Secretary of Education carries out *Olimpiadas del Conocimiento*, a series of competitions between all private and public schools. Individuals are split into two categories (grade 5 and grades 10-11), with competition happening within each of them in several stages. In 2016, the competition was carried out in August and there were over 83.000 students participating in the first stage between both categories, with the 3.500 best students moving on to the second stage. From here, 25 students per category were selected to participate in the semifinals and only five of them moved on to the finals, with a single winner within each group. Awards were given to all students who classified to the semifinals, but only the ten finalists received the biggest prizes (a full university scholarship for the grade 10-11 finalists). Even though the scholarship award could be quite important for low-income students, results do not have any relevance outside of the awards, unlike *Saber 11*. Also note that the chances of getting to the finals are of 0.01%, and the winners are typically students from elite private institutions. Therefore, these exams are not really high-stakes for public school students.

Using the same methods described above and administrative records from the 2015-2016 competitions, we find no discernible impact of the program on *Olimpiadas del Conocimiento*'s scores. Table 9 shows the results of a doubly robust estimation for grade 10 and grade 11 students, as well as a joint estimation. Effects are not statistically

different from zero on the students' rank or the standardized test scores.

Table 9: Effects on student's performance in *Olimpiadas del Conocimiento*

	Student's rank			Standardized test scores		
	(1) Grade 10	(2) Grade 11	(3) Joint	(4) Grade 10	(5) Grade 11	(6) Joint
SaberEs effect (β)	-0.271	-1.647	-0.992	-0.009	-0.062	-0.036
	(1.056)	(1.213)	(0.901)	(0.038)	(0.043)	(0.032)
Observations	35,852	31,592	67,444	35,852	31,592	67,444
Controls	YES	YES	YES	YES	YES	YES

Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). All specifications control for stratum, family income, parent's education, mobile phone ownership and student's working status. *p<.05; **p<.01; ***p<.001

The results imply that the mechanism that explains the greater access to higher education is not a motivational effect. Students were not encouraged to learn and accumulate human capital beyond the techniques they learned in the program, since their performance on *Olimpiadas del Conocimiento* was unaffected. Similar to the literature in labor economics that shows that task-specific human capital is an important source of individual wage growth (Gathmann & Schönberg, 2010; Gibbons & Waldman, 2004), we show that it is also associated with strong effects on access to tertiary education.

An additional mechanism through which *SaberEs* may have impacted access to tertiary education is higher access to financial aid and scholarships. In Colombia, *Saber 11* is used as a selection mechanism to access the main sources of financial aid and scholarships. Similar to Bernal and Penney (2019); Londoño-Vélez et al. (2020); Melguizo et al. (2016), we test the effects of *SaberEs* on such outcomes. Our outcomes of interest are access to regular public financial aid (ICETEX or Sapiencia), and access to *Ser Pilo Paga* (SPP) the main merit-based scholarship in the country at that time.

Since the program had a positive impact on the median and top-performing students' test scores, this may have made them pass the required threshold for financial aid and encouraged them to compete for such resources. Thus, we would expect to see positive effects on access to ICETEX or Sapiencia. As shown in Table 10, there are no significant effects on access to either one of the agencies.[46] This implies that access to these educational credit options does not explain our higher education results, so this particular mechanism can be ruled out.

[46]Results are the same when estimating on each of the agencies individually.

Table 10: Effects on access to financial aid and *Ser Pilo Paga*

	(1) Higher Education	(2) Financial Aid	(3) Ser Pilo Paga
SaberEs effect (β)	0.037***	0.005	0.010***
	(0.013)	(0.005)	(0.004)
Observations	35,484	35,484	35,484
Controls	YES	YES	YES
Mean Control 2015	0.677	0.0530	0.0464

Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

However, for a small elite group of individuals (slightly above the 90th percentile), merit-based financial aid could have been an explanation for their post-secondary enrollment. In Figure 4 we show that *SaberEs* has positive effects on scores above the SPP cut-off. Now, we test whether the program has a positive impact on access to SPP. Column 2 in Table 10 shows that the program increased access to this scholarship by 1 percentage point on average. This corresponds to a striking 21.6% relative effect from a baseline value of 4.64%. SPP was a very generous scholarship that allowed socioeconomically disadvantaged students to enroll in tertiary education, especially those at the top of the distribution of skills.

Using administrative records from SPP, we are able to see what type of program the winners of the scholarship enrolled in, which allows us to isolate the effects by short-cycle and professional programs. Table 11 shows the results of a doubly robust estimation of the effects of *SaberEs* on access to SPP based on the student's higher education program of choice. The results show that for this selective group of individuals, merit-based financial aid was an effective mechanism to increase access to both short-cycle and professional programs. Nonetheless, the effects are more pronounced on professional programs, where the point estimate associated with access to the SPP scholarship is 1.0 percentage points (22% increase in relative terms). These results prove that access to merit-based scholarships was an important mechanism through which *SaberEs* impacted higher education enrollment, at least for a small group of elite students.

Table 11: Effects on access to SPP by type of program

	(1) Short-cycle SPP	(2) Professional SPP
SaberEs effect (β)	0.001*	0.010**
	(0.000)	(0.004)
Observations	35,484	35,484
Controls	YES	YES
Mean Control 2015	0.0007	0.0457

Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

7 Conclusion

In this paper, we estimate the effect of a standardized test preparation program called *SaberEs* on students' test score performance, and access to tertiary education. Its main objective was to develop and strengthen preparation for standardized cognitive tests, and ultimately improve access to tertiary education for students graduating from public schools in Medellin, Colombia. We take advantage of the lack of universal coverage in the first year of the program and its timing to estimate a difference-in-differences (DiD) and staggered event study specifications. Furthermore, we exploit our access to multiple administrative records to provide evidence on the main mechanism through which the program affected access to tertiary education.

SaberEs positively affects students' performance in *Saber 11*, the high school exit exam in Colombia. On average, students' rank within their cohort in *Saber 11* increases between 2.22 and 2.96 depending on the specification (or between 0.066 and 0.104 standard deviations). These results represent a reduction in the rank's gap between treated and control schools of 22.9%. The positive effect on *Saber 11* is driven by the effects on math, science, and social studies scores. In addition, most of the effects of the program are concentrated at the top-half of the score distribution, including percentiles above the Ser Pilo Paga (SPP) cut-off —the main merit-based scholarship in Colombia. We provided several estimations to mitigate concerns associated with the identification of the program's effect. Conclusions were similar in all cases. We also provide several informal and formal tests for parallel trends, our main identification assumption.

In addition, we show that *SaberEs* had a substantial effect on access to higher education, the program's main policy target. Our estimates suggest that the program led to a relative increase in access to tertiary education of 5% from a baseline value of 67.7%. Such effects were concentrated in STEM short-cycle programs. Furthermore, the program increased graduation from short-cycle programs by 2.3 percentage points on average, relative to a baseline average of 15.2%.

We then study the potential mechanism through which the program increased access to tertiary education. First, we explore the accumulation of task-specific human capital. *SaberEs* is a type of program that is especially helpful when students take high-stakes exams. We show that the effects are concentrated in highly-competitive higher education institutions that, in addition to *Saber 11*, require their own standardized test as a selection mechanism for admission. This sheds light on the effective accumulation of task-specific human capital. Using a low-stakes exam specific to the city of Medellin, we rule out the possibility of motivational effects.

Finally, we show that the effects on the upper percentiles of the *Saber 11* score affected the probability of accessing SPP, thus affecting access to higher education. SPP was a very competitive nationwide scholarship that used a specific cut-off in *Saber 11* as a selection mechanism. We show that the program positively affects percentiles above that cut-off. Our results indicate that *SaberEs* increased access to this scholarship by 1.1 percentage points on average, from a baseline value of 4.64%.

Our study is the first to provide evidence on *SaberEs*' causal effects. Moreover, this is one of the few papers that have analyzed these types of policies for socioeconomically disadvantaged students in Latin America, aside from Gómez et al. (2020). In this context, our paper might help the debate on public policies positively affecting access to higher education (Avery, 2013; Buchmann, Condron, & Roscigno, 2010; C. Cornwell, Mustard, & Sridhar, 2006; C. M. Cornwell, Lee, & Mustard, 2005; Richburg-Hayes et al., 2009; Rosa, Bettinger, Carnoy, & Dantas, 2022). A limitation of our exercise is the absence of a cost-benefit analysis. Nonetheless, given the low cost of the program,[47] its effects on scores and access to higher education, and the large return to tertiary education in Colombia,[48] we expect that the net present value of benefits of *SaberEs* was large and positive.

[47]The program's average cost per student in 2016 was $132.9 USD, and $168.8 USD in 2017 (calculated using the December of 2016 exchange rate of $3009.53 USD per COP). This is especially low when compared to other policies like SPP, which had an average annual cost per student of $3386 USD, using the value of the lowest possible stipend (Laajaj et al., 2022). In other words, *SaberEs* costed between 4% and 5% of the total cost of SPP in 2016 and 2017, respectively.

[48]The world's average rate of return to an additional year of education is 8.8%, while Colombia's oscillates between 10%-14% (Psacharopoulos & Patrinos, 2018). In fact, when compared to Latin American countries, Colombia's returns to higher education are the highest (Ferreyra et al., 2017).

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, *72*(1), 1–19.

Angrist, J., & Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American economic review*, *99*(4), 1384–1414.

Avery, C. (2013). Evaluation of the college possible program: Results from a randomized controlled trial. *National Bureau of Economic Research*.

Bernal, G. L., & Penney, J. (2019). Scholarships and student effort: Evidence from colombia's ser pilo paga program. *Economics of Education Review*, *72*, 121–130.

Bond, T. N., Bulman, G., Li, X., & Smith, J. (2018). Updating human capital decisions: Evidence from sat score shocks and college applications. *Journal of Labor Economics*, *36*(3), 807–839.

Borusyak, K., & Jaravel, X. (2017). Revisiting event study designs. *Available at SSRN 2826228*.

Borusyak, K., Jaravel, X., & Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*.

Bruce, D. J., & Carruthers, C. K. (2014). Jackpot? the impact of lottery scholarships on enrollment in tennessee. *Journal of Urban Economics*, *81*, 30–44.

Brunello, G., & Kiss, D. (2022). Math scores in high stakes grades. *Economics of Education Review*, *87*, 102219.

Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow education, american style: Test preparation, the sat and college enrollment. *Social forces*, *89*(2), 435–461.

Byun, S.-y., Chung, H. J., & Baker, D. P. (2018). Global patterns of the use of shadow education: Student, family, and national influences. In *Research in the sociology of education.* Emerald Publishing Limited.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Cohn, E., Cohn, S., Balch, D. C., & Bradley Jr, J. (2004). Determinants of undergraduate gpas: Sat scores, high-school gpa and high-school rank. *Economics of education review*, *23*(6), 577–586.

Cornwell, C., Mustard, D. B., & Sridhar, D. J. (2006). The enrollment effects of merit-based financial aid: Evidence from georgia's hope program. *Journal of Labor Economics*, *24*(4), 761–786.

Cornwell, C. M., Lee, K. H., & Mustard, D. B. (2005). Student responses to merit scholarship retention rules. *Journal of Human Resources*, *40*(4), 895–917.

De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–96.

Dobbie, W., & Fryer, J., Roland G. (2011, July). Are high-quality schools enough to increase achievement among the poor? evidence from the harlem children's zone. *American Economic Journal: Applied Economics*, *3*(3), 158-87. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/app.3.3.158` doi: 10.1257/app.3.3.158

Duquennois, C. (2022, March). Fictional money, real costs: Impacts of financial salience on disadvantaged students. *American Economic Review*, *112*(3), 798-826. Retrieved

from https://www.aeaweb.org/articles?id=10.1257/aer.20201661 doi: 10.1257/aer.20201661

Ferreyra, M. M. (2021). Landscape of short-cycle programs in latin america and the caribbean. *The Fast Track to New Skills*, 33.

Ferreyra, M. M., Avitabile, C., Paz, F. H., Botero, J., & Urzúa, S. (2017). *At a crossroads: higher education in latin america and the caribbean.* World Bank Publications.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, *77*(3), 953–973.

Gathmann, C., & Schönberg, U. (2010). How general is human capital? a task-based approach. *Journal of Labor Economics*, *28*(1), 1–49.

Gibbons, R., & Waldman, M. (2004). Task-specific human capital. *American Economic Review*, *94*(2), 203–207.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: the role of effort on the test itself. *American Economic Review: Insights*, *1*(3), 291–308.

Gómez, S. C., Bernal, G. L., & Herrera, P. (2020). Test preparation and students' performance: The case of the colombian high school exit exam. *Cuadernos de Economía*, *39*(79), 31–72.

Goodman, J., Gurantz, O., & Smith, J. (2020, May). Take two! sat retaking and college enrollment gaps. *American Economic Journal: Economic Policy*, *12*(2), 115-58. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pol.20170503 doi: 10.1257/pol.20170503

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Gurantz, O., & Odle, T. K. (2020). The impact of merit aid on college choice and degree attainment: Reexamining florida's bright futures program. *Educational Evaluation and Policy Analysis*, 01623737211030489.

Hájek, J. (1971). Discussion of 'an essay on the logical foundations of survey sampling, part i', by d. basu. *Foundations of statistical inference*, 326.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, *64*(4), 605–654.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, *47*(260), 663–685.

Kinsler, J., & Pavan, R. (2015). The specificity of general human capital: Evidence from college major choice. *Journal of Labor Economics*, *33*(4), 933–972.

Laajaj, R., Moya, A., & Sánchez, F. (2022). Equality of opportunity and human capital accumulation: Motivational effect of a nationwide scholarship in colombia. *Journal of Development Economics*, *154*, 102754.

Londoño-Vélez, J., Rodríguez, C., & Sánchez, F. (2020). Upstream and downstream impacts of college merit-based financial aid for low-income students: Ser pilo paga in colombia. *American Economic Journal: Economic Policy*, *12*(2), 193–227.

Medellin Mayor's Office. (2016). Development plan 2016-2019 "medellín cuenta con vos". *Medellín: Alcaldía de Medellín*.

Melguizo, T., Sanchez, F., & Velasco, T. (2016). Credit for low-income students and

access to and academic performance in higher education in colombia: A regression discontinuity approach. *World development*, *80*, 61–77.

Ost, B. (2014). How do teachers improve? the relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, *6*(2), 127–51.

Page, L. C., & Scott-Clayton, J. (2016). Improving college access in the united states: Barriers and policy responses. *Economics of Education Review*, *51*, 4-22. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0272775716301248` doi: https://doi.org/10.1016/j.econedurev.2016.02.009

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: a decennial review of the global literature. *Education Economics*, *26*(5), 445–458.

Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *The Review of Economic Studies*.

Richburg-Hayes, L., Brock, T., LeBlanc, A., Paxson, C. H., Rouse, C. E., & Barrow, L. (2009). Rewarding persistence: Effects of a performance-based scholarship program for low-income parents. *Available at SSRN 1353360*.

Rosa, L., Bettinger, E., Carnoy, M., & Dantas, P. (2022). The effects of public high school subsidies on student test scores: The case of a full-day high school in pernambuco, brazil. *Economics of Education Review*, *87*, 102201.

Roth, J. (2022). Pre-test with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2022). *What's trending in difference-in-differences? a synthesis of the recent econometrics literature.*

Sant'Anna, P. H., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, *219*(1), 101–122.

Webber, D. A. (2016). Are college costs worth it? how ability, major, and debt affect the returns to schooling. *Economics of Education Review*, *53*, 296–310.

World Bank. (2018). *World bank education overview : Higher education (english)*. World Bank Group.

Zwier, D., Geven, S., & van de Werfhorst, H. G. (2020). Social inequality in shadow education: The role of high-stakes testing. *International Journal of Comparative Sociology*, *61*(6), 412–440.

A Data construction appendix

In this section, we outline our main approach to data construction. As such, we provide a detailed description of the restrictions we placed on our datasets, as well as the way in which we constructed our treatment and outcome variables. Additionally, we explain how data from different sources were matched. Table A.1 summarizes our datasets and their sources.

Table A.1: Datasets

Name	Source
Saber 11	Icfes
Reports on treated schools	Secretary of Education of the Mayor's Office of Medellin
SNIES	National Ministry of Education
Saber TyT	Icfes
Olimpiadas del Conocimiento	Secretary of Education of the Mayor's Office of Medellin
Beneficiaries of financial aid	ICETEX
Beneficiaries of financial aid	Sapiencia
Beneficiaries of *Ser Pilo Paga*	ICETEX

Ultimately, we will have a unique individual-level dataset with all the information on test scores, access to higher education, types of tertiary programs, access to financial aid and scholarships, and socioeconomic characteristics for approximately 147,600 students from public schools in the city of Medellin.

A.1 Saber 11

The initial dataset from *Saber 11* contains information on around 4,360,000 students countrywide between 2010 and 2017. The major reduction in sample size comes from the exclusion of students who were enrolled in schools outside of Medellin (which were not reasonable for inclusion in the control group, given that the program took place in Medellin alone), and students enrolled in private schools (private schools already offered their students preparatory courses so they are not suitable for comparisons). These two restrictions leave us with around 147,600 observations. That will be our full sample size for the staggered event studies, while for the 2x2 case (2015-2016) we work with a subset of 35,495 students.

Next, we clean the high school's name variable to remove any unwanted characters like extra spaces, hyphens, and question marks. This allows us to match with the Secretary of Education's records of treated schools, which were also cleaned and capitalized to match the *Saber 11* format. We then match our main sample with each of the company's records (*Los Tres Editores* and *Avancemos*), and create our treatment variable as a dummy that takes the value of 1 if the match was successful, and 0 otherwise. Notice that in this case, our control group is comprised of public schools who did not appear in the treatment records. Given that only 150 schools were treated, we validated this procedure by individually checking that all treated schools were correctly matched in the

sample. It is important to note that there was no overlap in treatment assignment (i.e., each company had a separate set of schools).

We additionally create a second treatment variable to indicate the second wave of the treatment. This was done in the same way, but matching with the 2017 records instead. In both cases, these variables were interacted with a "post" variable that takes the value of 1 if the year is 2016 or 2017, and 0 otherwise. This interaction will capture the effect in the simple DiD specification. Additionally, for the Callaway and Sant'Anna (2021) estimator, we construct a variable that indicates when a school was first treated; e.g., if school A was treated in 2016 and 2017, and school B entered the program in 2017, their respective values will be 2016 and 2017. It will be 0 for the control group.

After we create our treatment variables, we move on to define our main outcomes: *Saber 11*'s rank and standardized scores. The former is more robust to skewed distributions than the latter and is less arbitrary than using a logarithmic transformation. To calculate it we first sort the students' total scores in the test from lowest to highest and assign them a value (x_i) depending on their position (1 being the lowest and N the highest, with N being the number of i students in the cohort). Then, we rescale this value from 0 to 100 as:

$$Rank_i = \frac{x_i - min(x)}{max(x) - min(x)} * 100$$

where the minimum and maximum values of x are cohort-specific.

On the other hand, for the standardized scores, we take the total scores and apply a simple standardization as:

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

where the mean and standard deviations are also cohort-specific.

Finally, we proceed to define our controls. All our covariates are defined as dummies equal to 1 if the description is satisfied and 0 otherwise (see Table 1's notes for the most important definitions). In the case of missing values, we assume that not answering a specific question is an answer in and of itself, so we assign a value of 0 to the few observations where it happens. This guarantees that our sample size will be consistent across specifications.

A.2 Higher education (SNIES and Saber TyT)

To construct the higher education variables, we merge our main sample data at the individual-level with the SNIES data. SNIES collects all the information on the higher education system in Colombia. This system integrates a comprehensive dataset with information on tertiary education institutions, programs, and students enrolled in and graduated from the different programs. With this data, it is possible to identify the institution attended, the degree obtained, and the graduation year for all individuals.

The main *Saber 11* data described above was merged with the SNIES data using variables such as ID number, name and date of birth. The match was done annually, i.e., we took the entire *Saber 11* sample and merged it year by year with the SNIES dataset. For instance, each cohort was followed one, two, and three years after their high school

graduation (the 2015 cohort was matched with 2016, 2017 and 2018, and the 2016 cohort was matched with 2017, 2018 and 2019). We did it in this way to ensure that we had annual information for the main estimates, and that comparisons based on years post graduation were accurate. It is important to keep in mind that given the panel nature of the SNIES data, one student may appear each year for which we calculate the estimates (if they are still enrolled in the program). In addition, there are students who may be enrolled in professional and short-cycle programs at the same time. In the construction of our dataset we did not exclude these students from the sample. They will instead take a value of 1 in both the professional and short-cycle dummies.

Despite the fact that SNIES includes information on student's graduation, we took the *Saber TyT* data as a proxy for graduation from short-cycle programs since the SNIES data presents a slight delay in these variables. *Saber TyT* is a mandatory exit exam for all students who are close to graduation from short-cycle higher education. As in the case of SNIES mentioned above, we merged this individual-level dataset with our main sample in order to construct a dummy variable that takes the value of 1 if the student graduated from a short-cycle program (if matched with the *Saber TyT* data). The variables and criteria for matching are exactly the same as the ones used in the SNIES matching process mentioned above.

A.3 Olimpiadas del Conocimiento

Given that *Olimpiadas del Conocimiento* was also a standardized exam that we use to test the motivational effect of the program, its data processing follows a similar pattern to that of *Saber 11*. We begin by taking the full sample, which contains around 575,600 observations between 2011 and 2018, and keep students from public schools only. Additionally, we drop all grade 5 students who are also in the full sample. This leaves us with a total of 314,635 students, out of which there are around 67,400 individuals that constitute our 2x2 sample (2015-2016). After focusing on our relevant students, we proceed by cleaning the schools' names as we previously explained in order to match them with the Secretary of Education's records. As a result, we create our treatment variables.

In this case, we do not have a total score as in *Saber 11*, but our "score" variable is the number of correct answers in the test. To be consistent with what we did in the *Saber 11* process, we create two main outcome variables: student's rank and standardized scores. Their calculation follows the same logic as before.

Finally, *Olimpiadas del Conocimiento*'s dataset also contains socioeconomic characteristics, although not as many as *Saber 11*. In this case, we have access to family income, socioeconomic stratum, both parent's education, whether the student works, and whether the student has a mobile phone. Their definition follows the same process as with *Saber 11*, assigning missing values a value of 0.

A.4 Other data

In addition to the aforementioned matches, we constructed outcome variables from the last three datasets listed in Table A.1 and referred to in the data section (section 3). For all of these, we have information at the individual level with all the identifying variables that allow us to match the main sample of *Saber 11* to all the mentioned secondary

datasets. Specifically, we perform this merge using different pairing criteria that combine the coincidence of identity document, names and date of birth of the individuals between one dataset and another. Although the matching is done independently between *Saber 11* and each secondary dataset, we use the same algorithm to avoid affecting the probability of matching.

After we have matched the datasets to our main sample, we create the relevant outcome variables. All of them are dummy variables that take the value of 1 if a student was matched to the respective dataset. However, in the case of ICETEX and Sapiencia, the variable we ultimately use for estimation is whether the student appeared in any of them (results are the same when estimating separately).

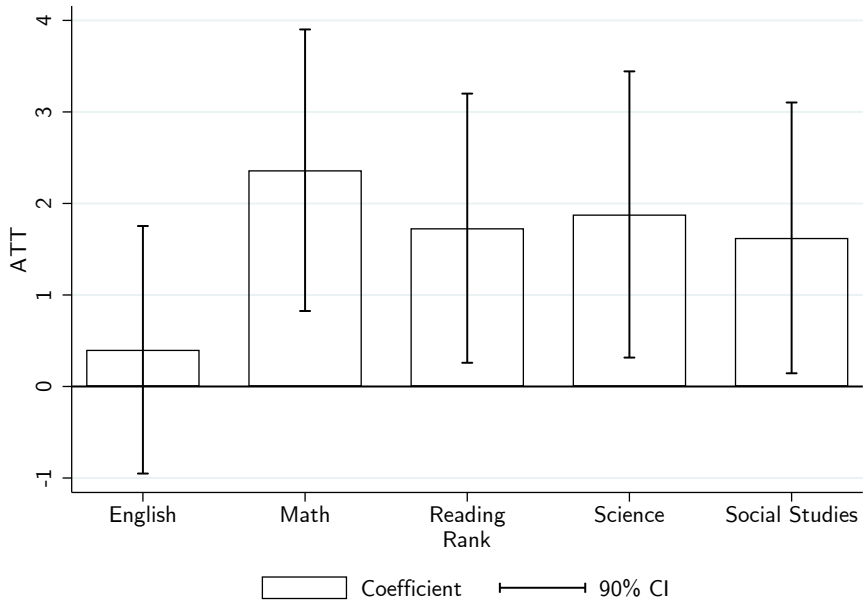
B Test scores appendix

Table B.1: Main results: Gap reductions

	(1) Student's rank	(2) Standardized test scores
SaberEs effect (β)	2.222**	0.074**
	(0.915)	(0.032)
Observations	35,484	35,484
Controls	YES	YES
Gap reduction with control group	22.9%	22.2%
Gap reduction with elite private	4.7%	3.4%
Gap reduction with elite public	9.1%	9.2%
Gap reduction with private	20.0%	17.8%

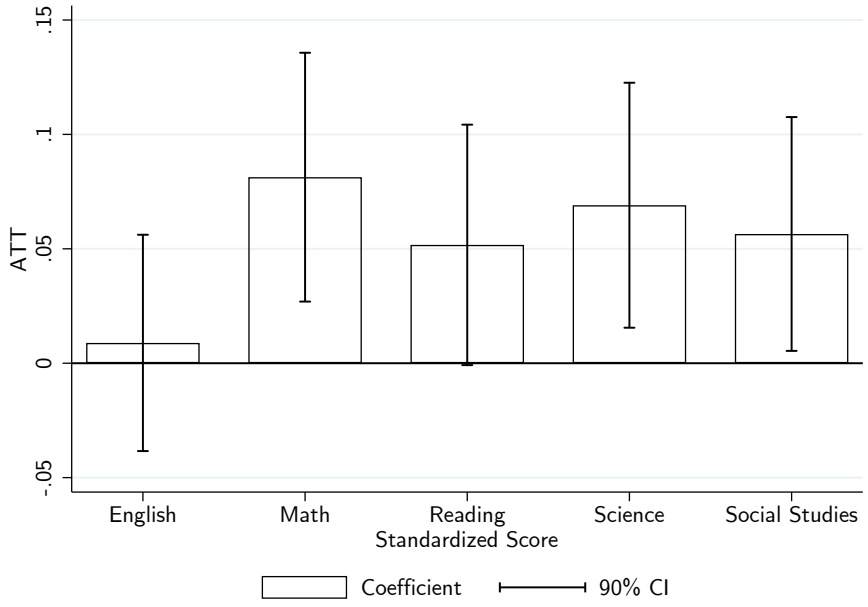
Notes: Standard errors clustered at the school level. The average rank and standardized score for the first gap reduction (control group) was calculated using our estimation sample (public schools only). As for the rest of the gap reductions, the pre-existing gaps over which we divide the estimated coefficient were calculated using the universe of schools in Medellin (public and private schools). We define elite private (public) schools as the ten private (public) schools with the best rank in 2015. In elite private schools, tuition costs are especially high (higher than the monthly minimum wage), which are unreachable for our sample. The reported coefficients come from the doubly robust estimation as in Sant'Anna and Zhao (2020). The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap. *p<.05; **p<.01; ***p<.001

Figure B.1: Effects on specific subject areas (rank)



Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap.

Figure B.2: Effects on specific subject areas (standardized score)



Notes: Standard errors clustered at the school level. Results come from a doubly robust estimation as in Sant'Anna and Zhao (2020). The controls used in the regressions are the ones in Panel D of Table 1, but we exclude NSE 1 to avoid the dummy trap.

Table B.2: 2010-2016 Results

	(1) Student's rank	(2) Standardized test scores
SaberEs effect (β)	2.384***	0.083***
	(0.615)	(0.021)
Gap reduction	24.6%	24.8%
Observations	53,297	53,297
Controls	YES	YES

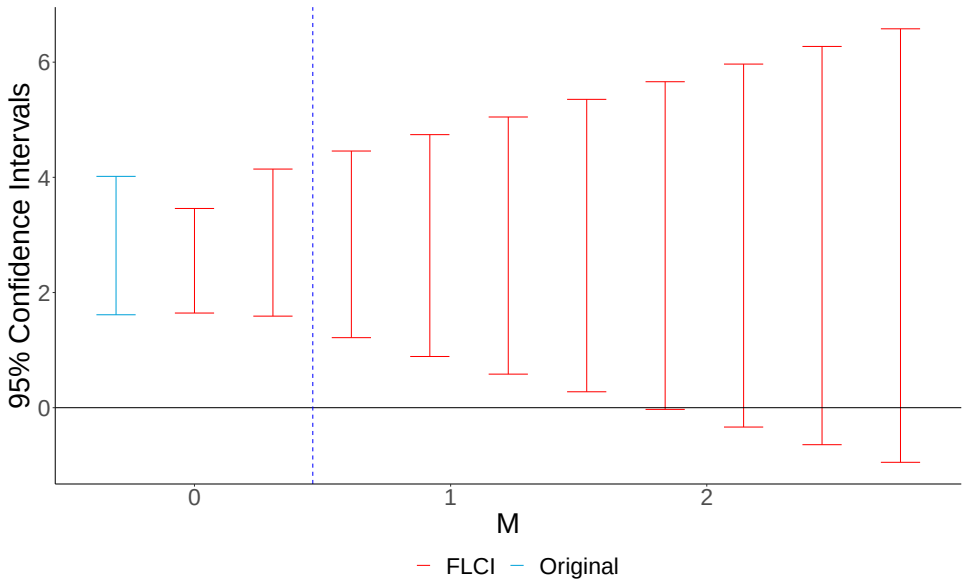
Notes: Standard errors clustered at the school level. Results come from a two-way fixed effects estimation that includes school and year fixed effects. *p<.05; **p<.01; ***p<.001

Table B.3: Power analysis: bias from hypothesized trend

	(1) Estimate	(2) Slope	(3) Likelihood ratio
Student's rank	3.715	0.462	0.009
Standardized test scores	0.131	0.016	0.009

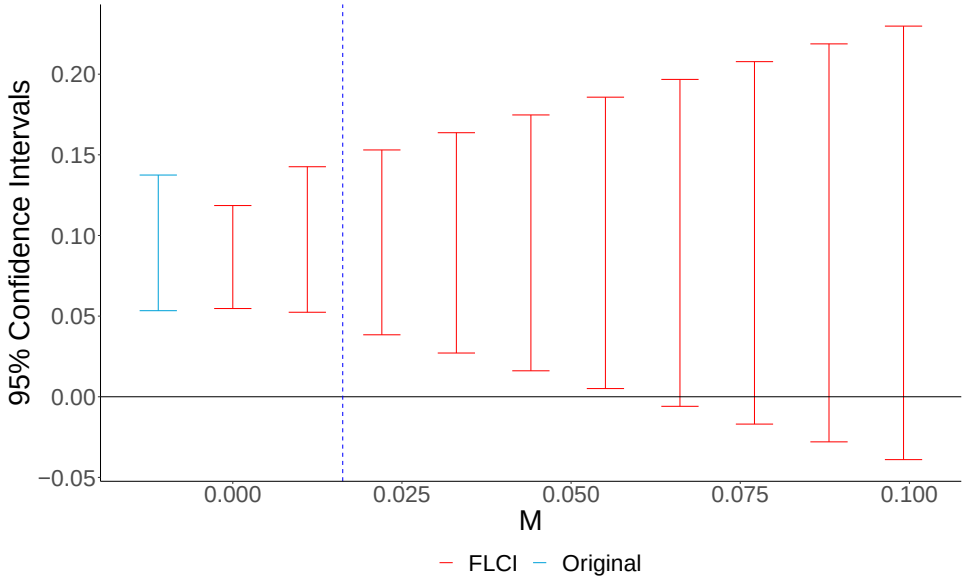
Notes: Column 1 displays the estimated "simple" coefficient from Table 3. Column 2 shows the pre-trend that has 50% power of being detected (hypothesized trend). Column 3 shows the likelihood ratio.

Figure B.3: Sensitivity analysis: student's rank



Notes: Based on Rambachan and Roth (2023). The dotted blue line represents the pre-trend that has 50% power of being detected shown in Table B.3.

Figure B.4: Sensitivity analysis: standardized test scores



Notes: Based on Rambachan and Roth (2023). The dotted blue line represents the pre-trend that has 50% power of being detected shown in Table B.3.


BANCO DE LA REPÚBLICA
COLOMBIA