SaberEs: test preparation program in Colombia

Christian Posso Estefanía Saravia Pablo Uribe*

April 8, 2022

Abstract

We study the effects of a citywide extracurricular program in Colombia that provided standardized test preparation and vocational guidance to senior year students from public schools. Particularly, we analyze how the program affected the performance of students in the national high school exit exam and their higher education enrollment rates. Using granular administrative data and following a 2x2 difference-in-differences design, we find that receiving the program significantly increased the students' average rank within the test at the median of the distribution. This effect corresponds to a 23% reduction in the pre-existing gap between control and treated students. When expanding our sample over time, results are robust to the recent econometric methods for dynamic difference-in-differences designs. Finally, we document a significant increase in the post-secondary enrollment rate up to three years after high school graduation of 7.1% from a baseline level of 52.4%. The effect is explained by the dynamics between technical and professional programs and corresponds to a striking 57% reduction in the pre-existing gap. Our results highlight how relatively simple programs can have real positive effects on academic outcomes of low-income students by providing them with more opportunities in life.

1 Introduction

Standardized test scores are of great importance to high school students who plan to pursue higher education studies given the ability of these tests to measure cognitive skills that in a sense may be predictive of college performance. Hence, the results obtained in these tests cause students to update their academic options in terms of selectivity, sector, and tuition value (Bond, Bulman, Li, & Smith, 2018).

However, students may have pre-established beliefs about their abilities and may decide not to take them, which is why making it compulsory to take tests such as the ACT[1] changes students' beliefs through the new information they receive (results), thus increasing enrollment in selective universities (Goodman, 2016). Nonetheless, these mandates can have unintended consequences like increasing high school dropout rates (Hemelt & Marcotte, 2013; Warren, Jenkins, & Kulick, 2006). In this case, students who fail the exams typically drop out of school, unless they are presented with other graduation opportunities, like a project-based pathway, which can improve post-secondary outcomes in education and employment (Lincove, Mata, & Cortes, 2022).

*Posso: Researcher *Banco de la República*. Saravia: *ICFES*. Uribe: *Universidad Eafit*. The opinions and possible errors contained in this document are the sole responsibility of the authors and do not commit *Banco de la República* or its Board of Directors.

[1]The ACT is a U.S. national standardized test designed to measure how prepared a person is for college. Along with the SAT, they are the two most widely used national tests.

Given that these alternatives are not used on a regular basis, importance is still placed on high school exit exams, so a lot of effort has been made to come up with programs to increase students' results. For instance, teacher incentive programs have been widely studied, especially among low-income contexts, with most studies finding a positive effect on test scores (Loyalka, Sylvia, Liu, Chu, & Shi, 2019; Mbiti, Muralidharan, et al., 2019; Mbiti, Romero, & Schipper, 2019; Muralidharan & Sundararaman, 2011) or reductions in the dropout rates (Gilligan, Karachiwalla, Kasirye, Lucas, & Neal, 2022). However, these programs have been somewhat controversial because of the perverse incentives they place on teachers, causing them to cheat on the standardized tests (Jacob & Levitt, 2003).

As pointed out by Glewwe, Ilias, and Kremer (2010), the positive effects of teacher incentive programs seem to be generated by test preparation activities. In line with this, it is common to find extracurricular strategies that increase students' academic opportunities, either to deepen their knowledge in certain areas or to prepare them for standardized tests. This has been framed in the literature as shadow education and its use has been evidenced around the world, although there is some variation in its application among countries. However, as several cross-country studies have shown, students with higher socioeconomic status participate more in these types of programs, especially in countries where testing has higher stakes, i.e., it is of greater importance for a student's future (Byun, Chung, & Baker, 2018; Zwier, Geven, & van de Werfhorst, 2020).

In the United States, the use of commercial test-preparation courses has become quite popular. Buchmann, Condron, and Roscigno (2010) find that students from privileged families tend to enroll in these courses to a greater extent and that higher SAT scores increase the likelihood of getting into the most selective colleges. Similarly, Park and Becks (2015) observe that more elite forms of preparation predict significantly higher SAT scores, although these courses appear to be especially useful for students of high socioeconomic status (Domingue & Briggs, 2009). Beyond this, coaching in preparation for college entrance exams (similar to the high school exit exams in terms of content), has been found to only increase scores modestly (Briggs, 2001).

We contribute to the literature on shadow education by providing evidence on the importance of these programs. Our study focuses on Colombia's second largest city, Medellin, where there are stark differences in the quality of education between public and private schools. In this context, we explore the change in test scores after the introduction of a policy that aimed to strengthen the cognitive skills of students in public schools.

We study *SaberEs*, a program implemented by the Mayor's Office in 2016 that provided an extracurricular test preparation course for students enrolled in public schools with additional vocational guidance and faculty training components. Leveraging the lack of universal coverage on the first year and the timing of the policy, we use a difference-in-differences (DiD) setup to estimate causal effects of the program on the high school exit exam (*Saber 11*) and access to tertiary education, and find evidence of significant positive effects on both the score and the students' rank within the test as well as a positive effect on higher education enrollment.

The latter is relevant given that a lot of the studies in this literature have focused on existing programs within educational institutions and a few of them have analyzed the effect of third-party programs, as is the case with College Possible, a program targeting low-income high school students. The program, in some ways similar to *SaberEs*,

provides an after-school curriculum including ACT and SAT preparation services, and college admissions and financing consultations. Spinney, Uekawa, and Campbell (2019) find that treated students have higher graduation rates than those in the control group; additionally, those who receive more hours of mentoring are more likely to earn a college degree in a timely manner (Howley & Uekawa, 2013). However, although Avery (2013) also finds better results on application and enrollment to selective colleges, it finds little evidence of effects on ACT performance. Our study contributes to this literature by finding evidence of these effects in a developing country for a similar program.

In Colombia, students take the *Saber 11* test in their senior year and it is also common for private companies to offer preparatory courses commonly referred to as *Preicfes*, which are mostly utilized by private schools. Although there are some studies specifying that human capital and parental income significantly explain the results in *Saber 11* (Tobón, Posada, & Ríos, 2009), and that studying full-time also has positive effects (Chica, Galvis, & Ramirez, 2011)[2], there are no studies on the effect of these programs on test results (short term) and on higher education (medium term).

The closest to the latter are studies on *Ser Pilo Paga* (SPP), a nation-wide scholarship that funded 10.000 students' entire undergraduate education per year as long as they were from low-income households and scored above the 90^{th} percentile in the *Saber 11*. In this case, Londoño-Vélez, Rodríguez, and Sánchez (2020) find that the scholarship increased university enrollment, while Bernal and Penney (2019) find that it increased test scores for qualifying students. In addition to these studies, Laajaj, Moya, and Sánchez (2022) find that *Ser Pilo Paga* had a motivational effect on students to accumulate human capital, which in turn increased test scores and university enrollment.

We complement this literature by studying whether a less costly strategy increased students' abilities and therefore improved test scores. Our results imply that simpler policies can have the same impact as more resource-demanding ones like full scholarships. We find that *SaberEs* had a statistically significant and positive effect in terms of the student's rank within the test (higher than what some of the previously mentioned studies found) that were mainly driven by the effects on the median student and not those at the top of the distribution. In addition, we also find that the program had a positive impact on higher education three years after the students graduate from high school. Initially, they enroll in technical programs but later transition to professional ones.

The rest of the paper is organized as follows. Section 2 gives the context of the program. Section 3 describes the data and presents summary statistics. Section 4 details our empirical strategy. Section 5 presents the impacts of the program, and Section 6 concludes.

2 Context

The Colombian secondary education system runs from grade 6 to 11 and ends with students taking a compulsory exam called *Saber 11*. Upon graduation, students can decide to transition to higher education, either to a technical or technological program (T&T) or to a professional one. Regardless of the type of program students are enrolled

[2]In Colombia's public schools, there are full-time (regular schedule) or part-time (mornings, afternoons and nights) students.

in, they are legally obliged to take the *Saber TyT* or *Saber Pro*[3] to be able to graduate from their institutions. After this, they can either enter the job market or continue their studies in postgraduate education.

2.1 Saber 11

The *Saber 11* high school exit exam is a compulsory test similar to the SAT in the United States, and administered by ICFES[4], the institution responsible for measuring the quality of education through standardized testing. It is a compulsory exam with compliance rates above 90% and is a good indicator of student's cognitive skills (Bernal & Penney, 2019). On average, around 500,0000 students take it every year in either March or August, with the vast majority of them taking it in the latter.

The exam has had several structural changes since it began, with the most recent one happening in 2014. Before that year, the exam was divided into 8 subject areas: math, Spanish language, biology, physics, chemistry, social studies, philosophy, and English. However, as of 2014, the test's structure was changed to make its results comparable with other tests administered by ICFES. In this sense, it was divided into five subject areas: math, science, reading skills, social studies and English. Each area's score ranges between 0 and 100, and the test's general score ranges between 0 and 500 points, calculated as a weighted average of the individual tests.

Owing to the importance of the *Saber 11*, ICFES offers a familiarization test to students who want to prepare for it that used to cost around $30 USD,[5] something that low-income students cannot easily afford (Bernal & Penney, 2019). Additionally, private companies also offer courses with simulation exams and test-oriented classes, although these are mainly used by private institutions. Students from public schools, who are usually from low-income families, tend to miss out on these opportunities.

2.2 Higher education

The higher education system in Colombia is comprised of public and private institutions that carry out admission processes each semester. In these processes, students apply to specific programs but are not limited to a single institution or program. *Saber 11* plays a central role in the admission processes of these institutions (Londoño-Vélez et al., 2020), with its score being required by most of them as a selection mechanism. There are two main types of programs that institutions offer: technical and technological, and professional degrees.[6]

In addition, institutions are required to comply with quality standards set by the Ministry of Education to be able to operate, and as a signaling mechanism, they can also apply for a High Quality Accreditation. This certificate splits the supply of post-secondary education into high-quality and low-quality institutions, since it proxies quality of education provision (Camacho, Messina, & Uribe, 2017).

[3] *Saber TyT* is taken by technical and technological students while *Saber Pro* is taken by professional students.

[4] It stands for *Instituto Colombiano para el Fomento de la Educación Superior* in spanish.

[5] Currently, the ICFES' website contains free resources that anyone can access at any moment.

[6] Technical and technological programs have a duration of 2 and 3 years, respectively. Professional programs span from 4 to 5 years.

As pointed out by Ferreyra (2021), higher education costs in Colombia are especially high when compared to other countries in the region, primarily because of private professional programs which are significantly more expensive than public ones. Even if both professional and T&T public programs are highly subsidized by the government, the cost of studying one of these programs in a public institution is higher than in other Latin American countries, but is still accessible to low-income students. Due to this, the biggest public universities in the country have really competitive admission processes where only the highest-achieving students manage to get admitted. In the private institutions, however, funding is the main channel through which low-income students can enroll, but even if scholarships are mainly taken by the highest-achieving individuals, there is a large educational credit market where people can finance their studies.[7] Yet, the majority of students in high-quality private institutions are high-income and high-achieving individuals, while low-income students typically sort into low-quality institutions.

2.3 SaberEs

In 2016, the Secretary of Education of the Mayor's Office of Medellín implemented a strategy called *SaberEs*. This initiative arose from the 2016-2019 Development Plan, 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 abilities that strengthen preparation for standardized tests such as *Saber 11*, which in turn allow students to aspire to university scholarships and prepare for their admission exams. To accomplish this, the Development Plan establishes the use of simulations as a familiarization and diagnostic tool for students, as well as competency training sessions to develop skills in the analysis and solution of the types of questions specific to these tests. Additionally, the strategy also has a component of teacher training, installed capacity in educational institutions and vocational guidance for senior students.[8]

Consequently, in the first year of the program, two companies were hired by the Secretary of Education to carry out the strategy in the city's public institutions. Each company was assigned a separate set of official schools whose selection was primarily based on their past score in the ICFES' standardized tests[9], so the selected institutions could not anticipate their inclusion in the program. With these lists, each company was responsible for implementing the strategy in their assigned schools from grades 8 to 11 (senior year in Colombia), although most of their resources were concentrated on the senior students.

Specifically, these companies rolled out the program between June and July of 2016 in the following way. First, they focused on teacher training for the five subject areas of *Saber 11*. Second, they trained the schools' principals and coordinators in pedagogical and methodological strategies; and finally, they conducted a simulation test that was later accompanied by feedback sessions with both students and teachers. In the case of

[7]Most of the market is controlled by ICETEX, a large public institution responsible for providing student loans.

[8]In the vocational guidance component, students take a vocational and occupational orientation test that suggests areas of study based on their measured abilities.

[9]The first company, *Los Tres Editores*, was assigned a total of 100 institutions, while the second company, *Avancemos*, was assigned 52. It was only in 2019 that all public institutions were covered.

the company with almost two thirds of the assigned institutions' population, a total of three simulation tests were performed to grade 11 students, each of them later reviewed in feedback sessions that were followed by teacher training conferences with different contents every time.

It did not take long until it was anecdotally claimed by the press and policymakers that the program had a positive impact, significantly improving the public schools' test scores in 2016.[10] Furthermore, the vocational guidance component was also acclaimed for providing students with enough information to take a wise decision about their post-secondary education.[11]

3 Data

We use administrative data from three main sources. First, we use data from ICFES, which contains information on all grade 11 test-takers in the country between 2010 and 2017 including their test scores, school characteristics and self-reported socioeconomic characteristics such as household goods, parent's education and stratum. This data set is then restricted to test-takers from the second semester applications of each year in the city of Medellin, and we drop private school observations to ensure valid comparisons.[12] The other source of data is the Mayor's Office of Medellin. Based on public contracts, we were able to identify 151 official schools that received the program in 2016. This is then matched with the previous data set to determine each student's treatment status based on the institution where they studied.

Finally, we use data from the Ministry of Education, specifically the National Higher Education Information System (SNIES by its acronym in Spanish). It includes detailed information on all tertiary education in the country at the student level, such as the type of program they are enrolled in and the institution's characteristics. We focus on professional, and technical and technological students who are enrolled in a higher education institution for any given semester between 2016 and 2019. In this case we are not able to include more pre-treatment periods and do the same type of dynamic analysis we do with *Saber 11* given that SNIES started in 2016, and prior to its introduction the higher education data were stored in another system called SPADIES.[13] Since the two data sets are fundamentally different from each other in the way information was collected, we decide to work exclusively with SNIES to ensure adequate estimations and to be able to see effects up to 2019.

As mentioned in the previous section, *Saber 11* had a structural change in 2014, thus making test scores uncomparable in our sample period. To overcome this challenge, we use the student's rank within his cohort as the variable of interest and rescale it from 0 to 100 (worst to best, respectively) following Laajaj et al. (2022). This overcomes the 2014 challenge by ensuring the comparison of results between years even after the difficulty or the structure change, and is more robust than using the score levels or standardized

[10] `https://telemedellin.tv/pruebas-saber-11/152207/`

[11] `https://www.elmundo.com/portal/vida/educacion/test_ayudo_a_definir_la_vocacion_de_jovenes.php#.YhP31OjMK3B`

[12] This is especially important given that the vast majority of private schools offer their students preparatory courses for the *Saber 11*.

[13] Even with the introduction of SNIES, data were still collected for SPADIES but only until 2018.

values. However, all the analysis was also done with the standardized test scores which can be found in the appendix.

Most of the analysis is done in what is commonly referred to as a 2x2 difference-in-differences setup to be able to obtain the most accurate estimates of the program's impact. Based on this setup, we consider the years 2015 and 2016 for the main results given that extending our timeline to 2017 would introduce dynamic effects, which combined with the staggered adoption nature of the treatment would impose additional concerns on the estimates. In this sense, the statistics shown in this section are for the 2x2 case, although we also extend our results to the dynamic case in future sections.

The summary statistics for the 2x2 sample are displayed in Table 1. Panel A shows the test scores in levels. Panel B displays overall access to higher education and specific access by type of program. 54% of students who graduated in 2015 and 2016 accessed any form of higher education within three years from graduation. Panel C shows the treatment variables; in this case, 66% of the institutions were treated by at least one of the companies, with 46% being treated by *Tres Editores* and 20% by *Avancemos* specifically. Finally, panel D contains all the socioeconomic covariates obtained from the *Saber 11* data set. 57% of the students were female and 10% had at least one parent with some tertiary education. Consistent to the sample being conformed by public schools, only 4% of students' households had a high stratum and 7% of them had a high income.

As a purely descriptive exercise, Figure 1 shows the distribution of the standardized general scores for 2015 and 2016. There is an increase in test scores in the year of treatment, although it seems to be concentrated along the median students and not at the tails. This descriptive analysis hints towards the presence of heterogeneous treatment effects along the outcome's distribution, which is more formally analyzed in Section 5.

Figure 1: Standardized General Score Density 2015-2016.

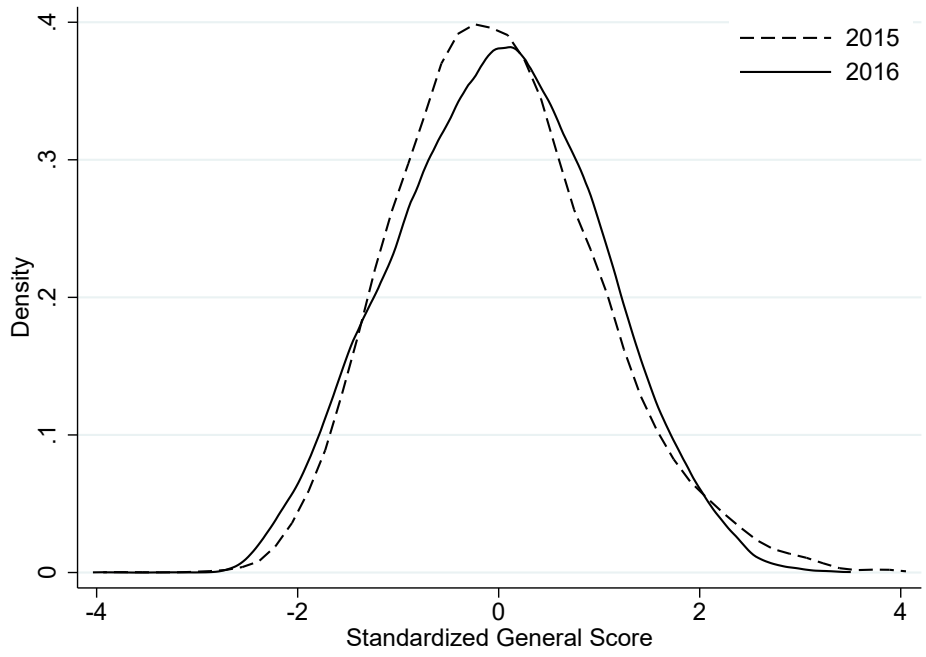


Table 1: Summary Statistics 2015-2016.

	Mean	SD	Min	Max
Panel A: Test Scores				
General	258.69	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				
Access to higher education	0.54	0.50	0	1
Access to T&T	0.29	0.45	0	1
Access to university	0.29	0.45	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 estimate the causal effect of *SaberEs* on the student's rank in *Saber 11*, we exploit the timing of the program and the lack of universal coverage using a difference-in-differences approach. 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:

$$Y_i = \alpha + \beta_0 \text{Treated}_i + \beta_1 \text{Post}_i + \beta_2 \text{TreatedxPost}_i + X_i'\delta + \varepsilon_i \tag{1}$$

where Y_i is the general rank of student i, *Treated* is a dummy variable indicating whether student i is part of a treated school, $Post_i$ takes a value of 1 if the student's test application year is 2016, and $TreatedxPost_i$ is their interaction. Finally, X'_i is a vector of controls that contains the socioeconomic covariates available from the *Saber 11* data, and ε_i are the standard errors, which are clustered at the school level. The coefficient of interest is β_2, which captures the ATT under the parallel trends assumption. We estimate the previous exercise with and without controls to check the robustness of the results after their inclusion, understanding that in that case we are implicitly assuming conditional parallel trends.

Additionally, we estimate a two-way fixed effects regression that takes into account the school and year fixed effects in order to control for unobserved heterogeneity. The regression is estimated as:

$$Y_i = \alpha + \theta_1 \text{TreatedxPost}_i + \psi_i + \gamma_i + \mu_i \tag{2}$$

where ψ_i and γ_i are the school and year fixed effects, respectively. Just as with the previous estimation, standard errors (μ_i) are clustered at the school level. Note that this specification does not control for covariates, since that would impose three additional assumptions that relate to the control's specific trends and to the homogeneity of treatment effects, so it could more likely lead to biased estimates of θ_1 (Sant'Anna & Zhao, 2020).

However, there are alternatives such as outcome regression (Heckman, Ichimura, & Todd, 1997) and inverse probability weighting (Abadie, 2005; Hájek, 1971; Horvitz & Thompson, 1952) that can estimate the ATT without imposing those additional assumptions, while also handling the inclusion of covariates. We use 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) estimator to take advantage of our large number of controls.

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.

In addition to the 2x2 estimations described above, we also analyze a dynamic specification in which we use observations from 2010 to 2017 to estimate the causal effect of the program on the student's general rank. 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 and estimate a reliable ATT in this dynamic setting, we use the Callaway and Sant'Anna (2021) estimator, which calculates all group-time specific ATT's. 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 to using an alternative specification proposed by Borusyak, Jaravel, and Spiess (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, Sant'Anna, Bilinski, & Poe, 2022). With this in mind, we focus mainly on the simple aggregation as recommended by Callaway and Sant'Anna (2021), and present the event study aggregation in the appendix.[14] Nonetheless, the parallel trends assumption is still critical in these procedures, and typical tests might not find significant pre-trends due to imprecise estimates. Because of this, we also calculate the possible bias from pre-testing following Roth (Forthcoming) and conduct a sensitivity analysis to check the robustness of the results when the parallel trends assumption might be violated as suggested by Rambachan and Roth (2021).

5 Results

5.1 Main results

In Table 2, Column 1, we report the difference-in-differences estimate without controls for the 2x2 case and find a statistically significant positive result. The coefficients in the table are directly interpretable as increases in the student's general rank in the *Saber 11*, so receiving the treatment corresponds to an effect of almost 3 ranks in the test. When including controls (Column 2), the effect slightly adjusts downward but stays at a significant 2.5 ranks.

When looking at more robust specifications, like outcome regression and inverse probability weighting with stabilized weights (Columns 4-5), the effect stays statistically significant and similar to the simple DiD estimates. However, the most important result is reported in Column 6, which displays a statistically significant doubly robust difference-in-differences estimate of 2.2 ranks. Also, Table A.1 in the appendix replicates the results using the standardized test scores instead of the rank and further shows the robustness of our results. In each case, to make the result more easily interpretable, the second row reports the coefficient in terms of the gap reduction. 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.[15] We find that the program generated a 22.8% reduction in the rank's gap between treated and untreated students.

These results are especially important when compared to what other studies have found using the *Ser Pilo Paga* scholarship. For instance, Laajaj et al. (2022) find that

[14] In order to avoid the issue of compositional changes in the event study aggregation, we present the results using balanced groups around the event time, as suggested by the authors.

[15] Basically, the calculation is performed as $ATT \,/\, E[Y_0 - Y_1 \mid t = 2015]$. The difference in means in this case is 9.67.

it increased the student's rank by 1.57 at the most, which was the case for students at the top of the distribution. Here, we are finding considerably higher ATT's across all specifications, and in the case of the DR estimation, almost an entire rank higher. On the other hand, Bernal and Penney (2019) find that SPP led to an increase of 0.09 test score standard deviations for the eligible students, and the effects were also concentrated at the top of the distribution. As reported in Table A.1, we find that *SaberEs* increased test scores by around 0.075 standard deviations, which is a similar effect to what they found. Yet in this case it is not driven by the top-performers.

Table 2: Main Results.

	(1) DiD	(2) DiD	(3) TWFE	(4) OR	(5) IPW	(6) DR
ATT	2.943***	2.538***	1.851**	2.196**	2.666***	2.207**
	(0.976)	(0.886)	(0.826)	(0.917)	(1.032)	(0.916)
Gap reduction	30.4%	26.2%	19.1%	22.7%	27.6%	22.8%
Observations	35,501	35,490	35,501	35,490	35,490	35,490
Controls	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 without controls, Difference-in-Differences with controls, Two-Way Fixed Effects without controls, Outcome Regression, Inverse Probability Weighting with stabilized weights, and Improved Doubly Robust Difference-in-Differences for repeated cross-section. *p<.05; **p<.01; ***p<.001

The results are even more impressive when looking at the costs of both interventions. As pointed out by Laajaj et al. (2022), SPP had an average cost per student of $2860 for an academic year without including the value of the stipend (1-4 monthly minimum wages per semester). The most conservative calculation, considering the lowest possible value of the stipend, yields an average cost of $3386 per student, which implies a total of over 33 million USD when considering the 10,000 students aimed to be covered by the program. In the case of *SaberEs*, the total cost for the first year of the intervention was 1.55 million USD[16], which represents only 4.6% of the total value of SPP. This suggests that a much cheaper policy is able to have comparable results to a significantly more expensive one, thus highlighting its cost-effectiveness.[17]

[16]This value was obtained by a back of the envelope calculation in which we converted the cost of 2016 in Colombian Pesos (COP) to 2015 COP, and then used the December 31^{st} exchange rate reported by the Central Bank to obtain a comparable value in USD.

[17]Although a point could be made that *SaberEs* was strictly oriented towards preparation for the tests and would have bigger effects per se, SPP was only granted to students who had a high relative score in the test so it had a direct positive incentive to perform well in the *Saber 11*.

5.2 Dynamic results and robustness

In this subsection, we present the estimates from the dynamic specifications that extend the sample from 2010 to 2017, as can be seen in Table 3. The first column reports the estimator proposed by Callaway and Sant'Anna (2021) and the second column the one proposed by Borusyak et al. (2021). Overall, the coefficients point to a similar conclusion -highlighting the robustness of our results- and are slightly bigger than the ones obtained in the 2x2 case. In terms of the gap reduction, the program reduced it by around 30 to 40%, given that it had a positive and significant effect of over 3 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, since the reduction of the gap is between 30 and 40%. This is presented in Table A.2 in the appendix.

Table 3: Dynamic Results.

	(1) CS	(2) BJS
ATT	3.704***	2.709***
	(0.785)	(0.497)
Gap reduction	39.3%	28.7%
Observations	147,675	147,573

Notes: Standard errors clustered at the school level. Column 1 displays the "simple" aggregation from the Callaway and Sant'Anna (2021). Column 2 displays the estimator based on Borusyak et al. (2021). *p<.05; **p<.01; ***p<.001

Furthermore, the results of the event study aggregation with balanced groups around the event time are consistent with what we have already found. These are presented in Figure A.2 and Figure A.3 in the appendix. 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. However, in order to further analyze the pre-trends we conduct a power and sensitivity analysis.

First, we check if the pre-treatment trends are parallel with a formal test as suggested by Roth (Forthcoming), as can be seen in Table A.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 (2021),

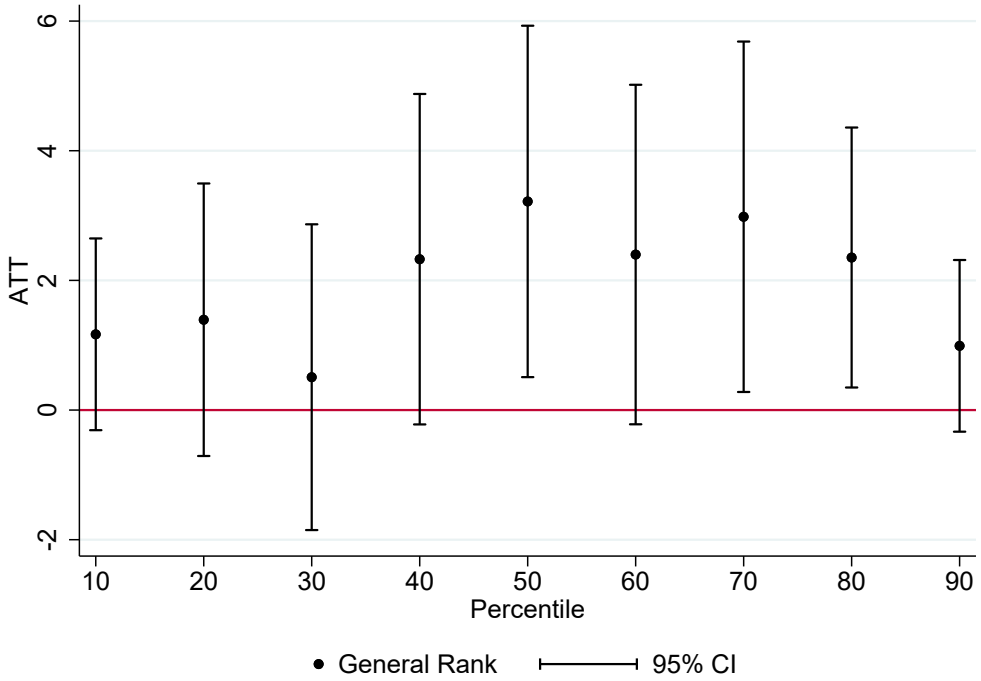
where we estimate a 95% confidence set for the general rank to consider its robustness to some degree (M) of deviation from the parallel trends assumption. Specifically, we check for linear (M=0) and non-linear (M>0) deviations. Figure A.4 reports the confidence set that results from this estimation. 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.

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 A.3 and Figure A.5), we arrive at the same conclusions.

5.3 Heterogeneous effects on the rank'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, Fortin, & Lemieux, 2009) to examine what is happening at the unconditional quantiles. This will allow us to determine whether the effects are concentrated among high- or low-performing students. Figure 2 shows the results of the RIF regressions over the deciles of the test rank distribution.

Figure 2: Effects by student's rank deciles (2015-2016).



Unlike the SPP studies mentioned before, the effects here are not concentrated in the

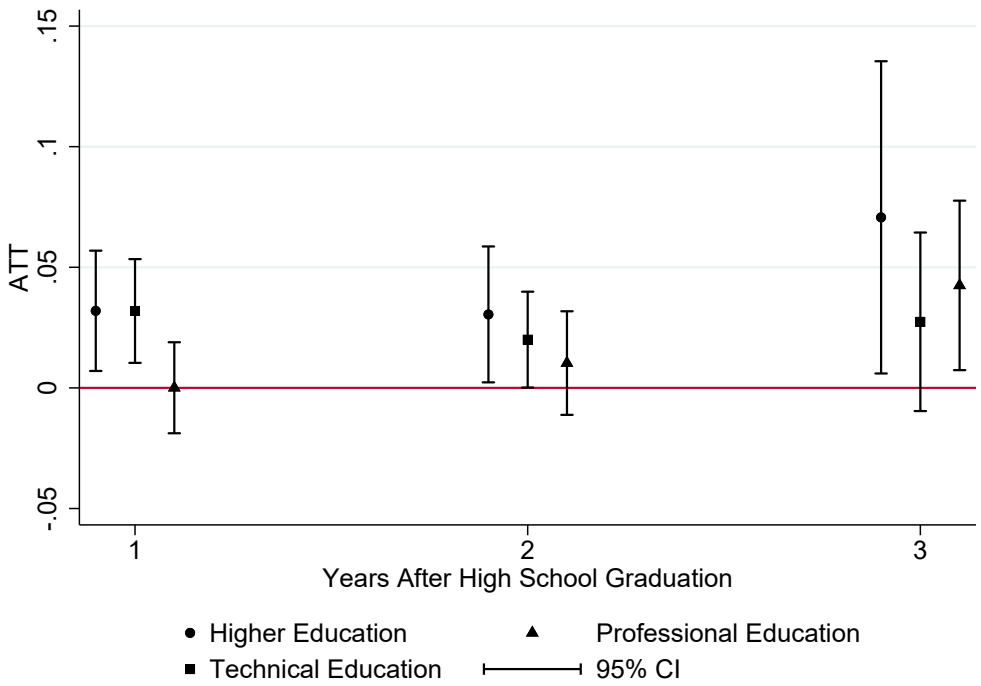
9^{th} decile (90^{th} percentile). In fact, effects are statistically significant at the 5^{th}, 7^{th} and 8^{th} deciles, with the highest one evidenced at the median corresponding to an effect of over 3 ranks. The results using the standardized test scores are displayed in Figure A.1 in the appendix and are consistent with the fact that our results are being driven by effects on the median student.

5.4 Higher education results

In order to look into the real effects of the program, we estimate the impact of *SaberEs* on access to higher education. In this sense, we estimate a doubly robust difference-in-differences regression as in Sant'Anna and Zhao (2020). Our main variables of interest are access to higher education, access to a technical or technological program, and access to a professional program. We split each of the outcomes into three groups to observe the dynamics of access to post-secondary education over time. As such, we look for results one, two, and three years after students graduate from high school. Figure 3 displays these results.

It is evident that the program had a positive impact on the immediate access to higher education the first year after graduation from high school, driven primarily by greater access to technical and technological programs. This overall effect persists over time. However, three years after graduation, access to professional programs is what maintains the positive effect on overall access. This implies that students are transitioning from technical to professional programs after they graduate from the former.[18]

Figure 3: Effects by student's rank deciles (2015-2016).



[18]This is consistent with the duration of technical programs (2 years).

These results are impressive for such a low-cost program, but are in line with its aim. The vocational guidance component appears to have been successful in providing students with enough information to actually enroll in higher education. This is likely due to the vocational and occupational orientation test they took, which highlighted their abilities and gave them a list of occupational field options where they were most likely to succeed. This, combined with their better results in *Saber 11* seems to have updated their beliefs and academic options, therefore increasing the overall access to post-secondary education by around 7% three years after their graduation from high school, from a baseline value of 52.4%. This effect is equivalent to a 57% reduction in the pre-existing access gap between control and treated students.

6 Conclusion

In spite of the considerable available evidence on the importance of standardized test performance for a student's future, not enough attention has been placed on the role of preparatory courses. Most of this evidence is concentrated in the United States, specifically on one particular program that targets low-income students, yet results from these studies are rather mixed. Therefore, not much is known about these programs' effectiveness in developing countries. This paper contributes to filling this gap.

We study the case of a program implemented in Medellin, Colombia, that offered a preparatory course to public schools throughout the city in an effort to increase their high school exit exam's test scores and reduce the gap between private and public institutions in the city. In this context, we take advantage of granular administrative data to identify the causal effect of the program on students' performance. To do this, we use recent econometric methodologies such as doubly robust estimators and other difference-in-differences estimators for dynamic settings.

Overall, we find that the program had a positive effect of more than 2 points on the average student's rank within the test and was concentrated around the median student, which is equivalent to a 23% reduction in the pre-existing gap. Additionally, using a dynamic setting that introduces staggered treatment adoption we find that the effect is slightly higher and translates into a 30% reduction in the gap. These results are especially important since other studies that have analyzed a full-scholarship program (considerably more costly) for low-income students in Colombia have found similar effects, even though they are concentrated at the top of the distribution (high-achieving students). This shows that our analyzed program had a real positive impact on students that could potentially translate to longer-term outcomes given the importance of test results on higher education institutions' admission processes and financial aid availability in the country.

We posit that simple programs have the potential to be as effective as more resource-demanding ones in their ability to impact certain outcomes. As such, our results are of great relevance to policy makers, who typically face budget constraints and are looking for the most cost-effective strategies to increase social welfare. Since the program started in 2016, long-term outcomes such as tertiary education graduation rates and labor market outcomes will only start to become available in the next years, so those issues should be explored by future research.

References

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

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*.

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from nels: 88. *Chance*, *14*(1), 10–18.

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.

Camacho, A., Messina, J., & Uribe, J. (2017). The expansion of higher education in colombia: Bad students or bad programs? *Documento CEDE*(2017-13).

Chica, S., Galvis, D., & Ramirez, A. (2011). Determinantes del rendimiento académico en colombia: Pruebas icfes saber 11, 2009 (academic performance determinants in colombia: Icfes saber 11, 2009 exam). *Center for Research in Economics and Finance (CIEF), Working Papers*(11-5).

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

Domingue, B., & Briggs, D. C. (2009). Using linear regression and propensity score matching to estimate the effect of coaching on the sat. *Multiple Linear Regression Viewpoints*, *35*(1), 12–29.

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

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

Gilligan, D. O., Karachiwalla, N., Kasirye, I., Lucas, A. M., & Neal, D. (2022). Educator incentives and educational triage in rural primary schools. *Journal of Human Resources*, *57*(1), 79–111.

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

Goodman, S. (2016). Learning from the test: Raising selective college enrollment by

providing information. *Review of Economics and Statistics*, *98*(4), 671–684.
Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.
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.
Hemelt, S. W., & Marcotte, D. E. (2013). High school exit exams and dropout in an era of increased accountability. *Journal of Policy Analysis and Management*, *32*(2), 323–349.
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.
Howley, C., & Uekawa, K. (2013). Evaluation of college possible postsecondary outcomes, 2007-2012. *ICF International*.
Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*(3), 843–877.
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.
Lincove, J. A., Mata, C., & Cortes, K. E. (2022). *A bridge to graduation: Post-secondary effects of an alternative pathway for students who fail high school exit exams* (Tech. Rep.). 53113 Bonn, Germany: Institute of Labor Economics (IZA).
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.
Loyalka, P., Sylvia, S., Liu, C., Chu, J., & Shi, Y. (2019). Pay by design: Teacher performance pay design and the distribution of student achievement. *Journal of Labor Economics*, *37*(3), 621–662.
Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, incentives, and complementarities in education: Experimental evidence from tanzania. *The Quarterly Journal of Economics*, *134*(3), 1627–1673.
Mbiti, I., Romero, M., & Schipper, Y. (2019). *Designing effective teacher performance pay programs: experimental evidence from tanzania* (Tech. Rep.). NBER: National Bureau of Economic Research.
Medellin Mayor's Office. (2016). Development plan 2016-2019 "medellín cuenta con vos". *Medellín: Alcaldía de Medellín*.
Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. *Journal of political Economy*, *119*(1), 39–77.
Park, J. J., & Becks, A. H. (2015). Who benefits from sat prep?: An examination of high school context and race/ethnicity. *The Review of Higher Education*, *39*(1), 1–23.
Rambachan, A., & Roth, J. (2021). An honest approach to parallel trends. *Working paper*.
Roth, J. (Forthcoming). 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.

Spinney, S., Uekawa, K., & Campbell, J. (2019). 2015 college possible: Closing the achievement gap for low-income students. i3 national development study. *Grantee Submission*.

Tobón, D., Posada, H. M., & Ríos, P. (2009). Determinants of the performance of the schools in medellin in the high-school graduation-year test (icfes). *Cuadernos de Administración*, *22*(38), 311–333.

Warren, J. R., Jenkins, K. N., & Kulick, R. B. (2006). High school exit examinations and state-level completion and ged rates, 1975 through 2002. *Educational Evaluation and Policy Analysis*, *28*(2), 131–152.

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 Appendix

Table A.1: Main Results: Standardized Score.

	(1) DiD	(2) DiD	(3) TWFE	(4) OR	(5) IPW	(6) DR
ATT	0.103***	0.088***	0.066**	0.073**	0.091**	0.074**
	(0.034)	(0.030)	(0.029)	(0.032)	(0.036)	(0.032)
Gap reduction	30.9%	26.4%	19.7%	21.8%	27.2%	22.1%
Observations	35,501	35,490	35,501	35,490	35,490	35,490
Controls	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 without controls, Difference-in-Differences with controls, Two-Way Fixed Effects without controls, Outcome Regression, Inverse Probability Weighting with stabilized weights, and Doubly Robust Difference-in-Differences for repeated cross-section. *p<.05; **p<.01; ***p<.001

Figure A.1: Effects by student's standardized test score deciles (2015-2016).

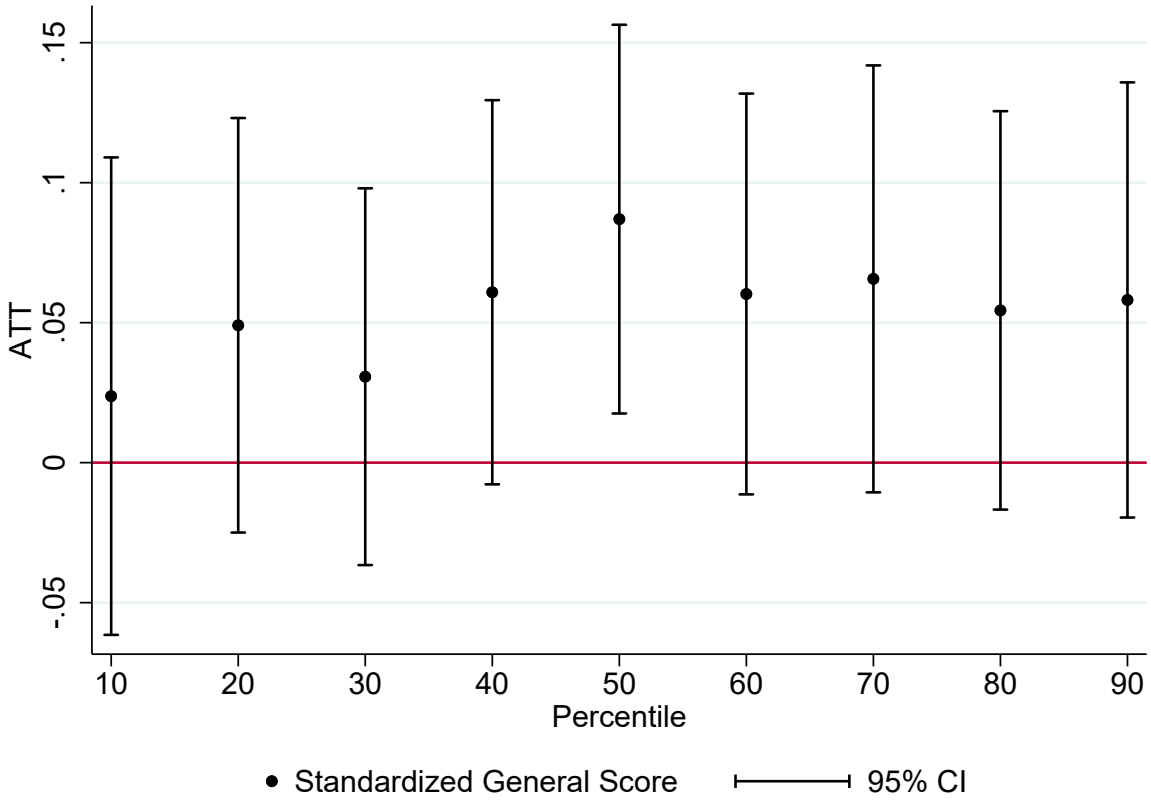
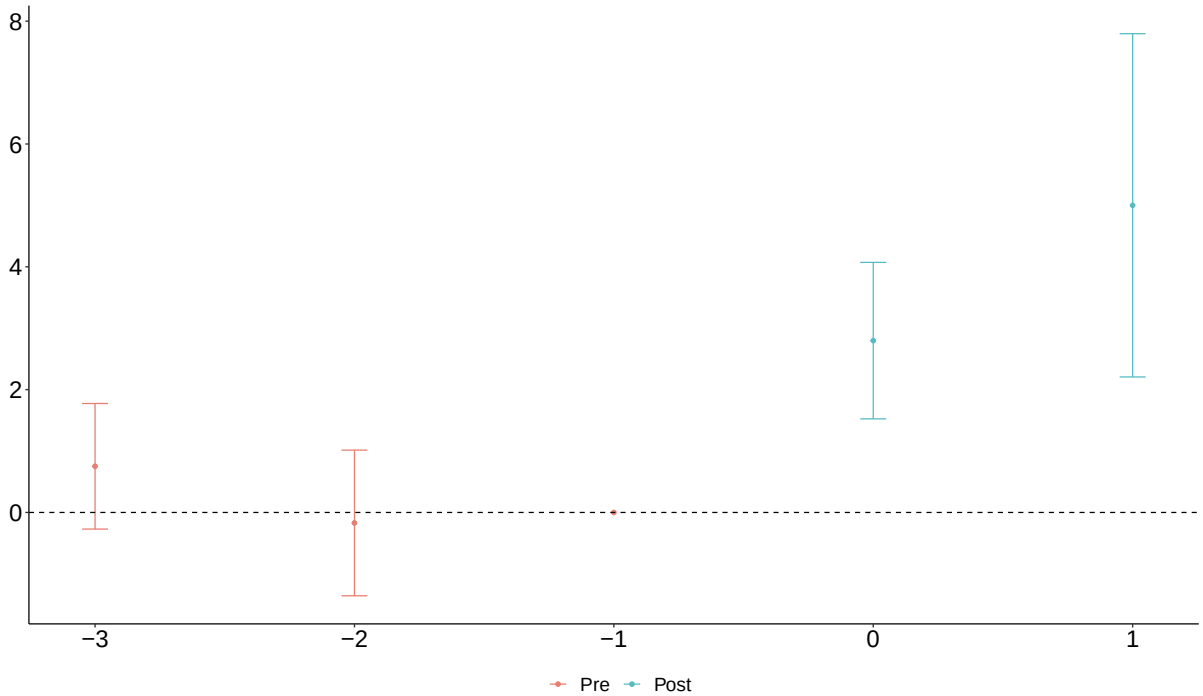


Table A.2: Dynamic Results: Standardized Score.

	(1) CS	(2) BJS
ATT	0.131***	0.099***
	(0.028)	(0.016)
Gap reduction	40.3%	30.4%
Observations	147,675	147,573

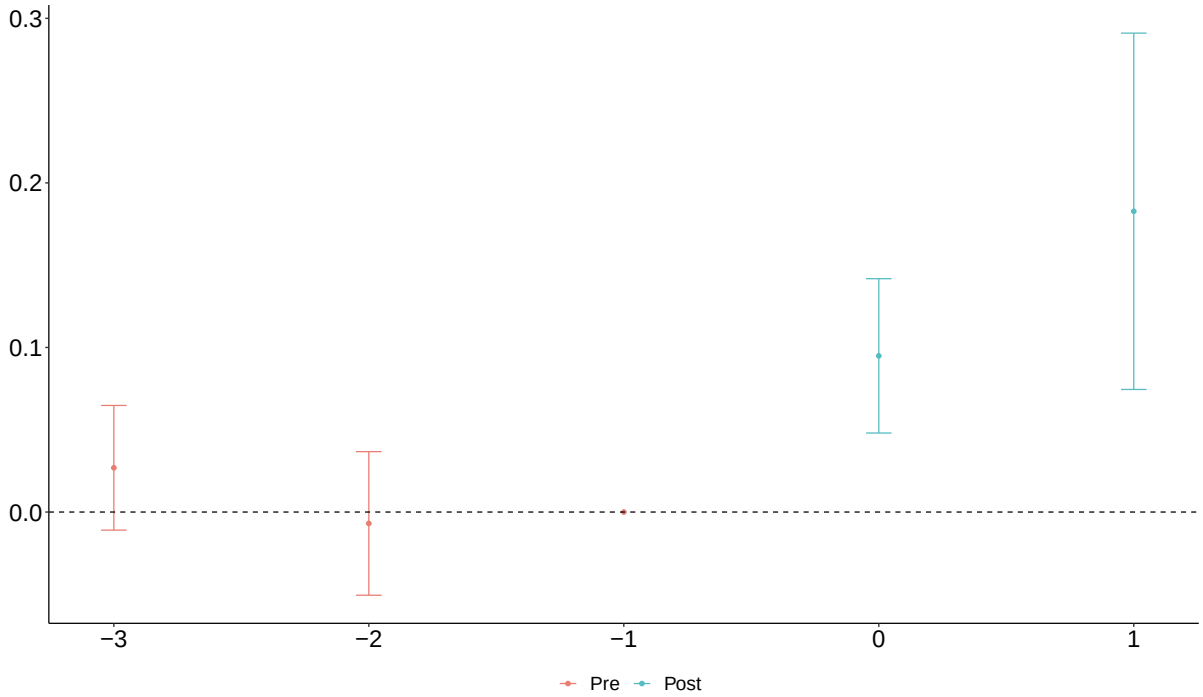
Notes: Standard errors clustered at the school level. Column 1 displays the "simple" aggregation from the Callaway and Sant'Anna (2021). Column 2 displays the estimator based on Borusyak et al. (2021). *p<.05; **p<.01; ***p<.001

Figure A.2: Average Effect on Student's Rank by Length of Exposure.



Notes: Estimation based on Callaway and Sant'Anna (2021).

Figure A.3: Average Effect on Student's Standardized Scores by Length of Exposure.



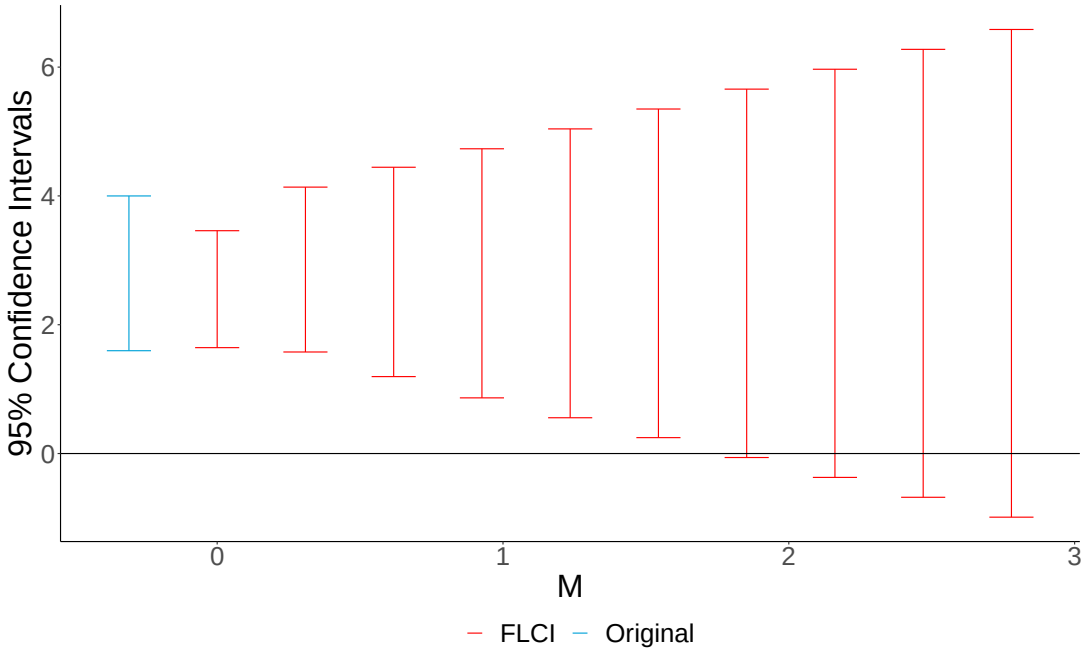
Notes: Estimation based on Callaway and Sant'Anna (2021).

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

	(1) Estimate	(2) Slope	(3) Likelihood ratio
General Rank	3.704	0.462	0.010
Standardized General Score	0.131	0.016	0.009

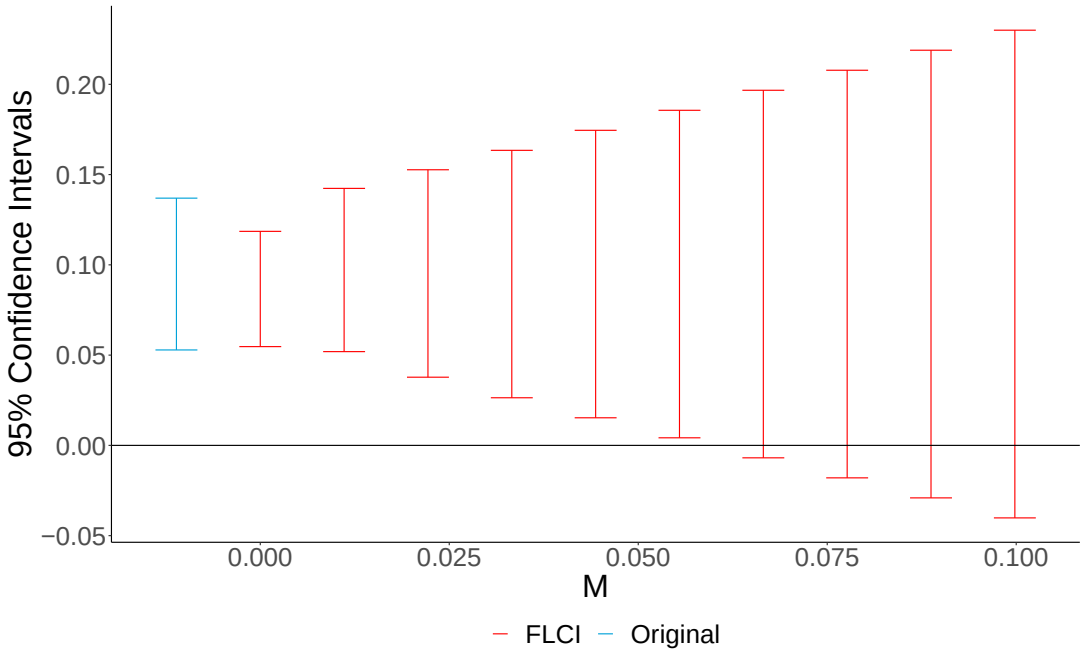
Notes: Column 1 displays the estimated "simple" coefficient from Table 3 and Table A.2. Column 2 shows the pre-trend that has 50% power of being detected (hypothesized trend). Column 3 shows the likelihood ratio.

Figure A.4: Sensitivity analysis: general rank



Notes: Based on Rambachan and Roth (2021).

Figure A.5: Sensitivity analysis: standardized general score



Notes: Based on Rambachan and Roth (2021).