Penney, Jeffrey; Lehrer, Steven F.; Bernal, Gloria L.; Reyes, Luis Carlos

Working Paper

Do opportunities for low-income students at top colleges promote academic success? Evidence from Colombia's Ser Pilo Paga program

Working Paper Series, No. 64

Provided in Cooperation with:
Canadian Labour Economics Forum (CLEF), University of Waterloo

Suggested Citation: Penney, Jeffrey; Lehrer, Steven F.; Bernal, Gloria L.; Reyes, Luis Carlos (2023) : Do opportunities for low-income students at top colleges promote academic success? Evidence from Colombia's Ser Pilo Paga program, Working Paper Series, No. 64, University of Waterloo, Canadian Labour Economics Forum (CLEF), Waterloo

This Version is available at:
https://hdl.handle.net/10419/279563

Canadian Labour Economics Forum

WORKING PAPER SERIES

Do Opportunities for Low-Income Students at Top Colleges Promote Academic Success? Evidence from Colombia's Ser Pilo Paga Program

Jeffrey Penney (University of Alberta)
Steven F. Lehrer (Queen's University)
Gloria L. Bernal (Pontifica Universidad Javeriana)
Luis Carlos Reyes (DIAN)

CLEF WP #64

Do Opportunities for Low–Income Students at Top Colleges Promote Academic Success?: Evidence from Colombia's Ser Pilo Paga program

Jeffrey Penney[1]
Department of Economics, University of Alberta
dr.jeffrey.penney@gmail.com

Steven F. Lehrer
Department of Economics, Queen's University and NBER
lehrers@queensu.ca

Gloria L. Bernal
Department of Economics, Pontificia Universidad Javeriana
gbernal@javeriana.edu.co

Luis Carlos Reyes
DIAN

version October 16, 2023

In 2014, the government of Colombia launched a unique means-tested and merit-based scholarship program called Ser Pilo Paga. We examine the effects of this scholarship on student performance on the country's university exit exam and other educational outcomes. Exploiting thresholds for socioeconomic status and test score performance on the high school exit exam for Ser Pilo Paga eligibility using a multi-score fuzzy regression discontinuity design, we find that the scholarship's effects on test scores to be quite limited in most cases. However, recipients of the scholarship do exhibit significant differences in enrollment, persistence, and test taking behaviour compared to non-recipients.

Keywords: Colombia, Saber Pro, scholarships, Ser Pilo Paga, test scores

JEL Classification: I22, I23, I28

[1]Acknowledgements: Penney and Lehrer acknowledge research support from SSHRC.

1. Introduction

Policymakers and higher education leaders of selective colleges and universities are increasingly stating a commitment to increase the enrollment of students that both excel academically and are overcoming challenging socioeconomic circumstances. Motivating this attention are beliefs that recruiting these students would first produce a more economically diverse student body on campus, and then subsequently increase the economic mobility and degree attainment for these students themselves. While research on low-income, high-achieving students has grown substantially in the past decade, it has primarily focused on their college application and enrollment decisions, with few studies examining whether public programs that provide a form of merit-based aid for students in need would influence outcomes after enrollment such as academic achievement, degree attainment, and early labour market outcomes.

Despite the backing of decisionmakers in higher education and others, it is evident that there are substantial challenges in attempting to accomplish the goal of increasing enrollment of high-achieving, low-income students at elite colleges;[2] in addition to funding concerns, any such policy would change the incentives faced by students, their parents, schools as well as the colleges and universities in operation. One government that attempted to successfully implement such a policy is the government of Colombia, who surprised the higher education sector in 2014 by announcing the creation of a scholarship called Ser Pilo Paga ("SPP"), a merit-based financial aid program that operates as a forgivable loan covering the full cost of tuition as well as providing a stipend for living expenses for each eligible recipient for the duration of an undergraduate degree program. To qualify for SPP, students needed to both score in the top 9% of test takers on the country's standardized high school exit examination, the Saber 11, and be at or below a given threshold of a socioeconomic index score ("Sisben") as measured by the government. One important feature of the SPP program was that the funds awarded came in the form of a loan which would be forgiven upon graduation, providing

[2]Early research using U.S. data including Pallais and Turner (2006) and Hoxby and Avery (2013) document that most very high-achieving students from low-income families do not even apply to a selective college or university. Subsequent research summarized in Lovenheim and Smith (2022) examined reasons for these gaps, focusing on either information barriers, social mismatch, credit constraints or complexity costs associated with the applications process.

students with a strong financial incentive to complete their degrees.[3]

In this paper, we examine the impact of SPP on both student academic outcomes and college admission decisions using administrative data from the government of Colombia's education and socioeconomic agencies for the 2014 and 2015 cohorts of graduating high school students. Specifically, we estimate the causal effect of SPP receipt on the performance on the country's standardized university exit exam, the Saber Pro, as well as other student outcomes such as choice of major, whether a student attends an elite university, whether a student attends a high cost program, student enrollment and persistence as measured by whether they ultimately write the university exit exam (which requires students to have obtained at least 75% of the necessary credits for graduating before being eligible to take), and the length of time from writing the Saber 11 exam to writing the Saber Pro exam. In addition, we pay particular attention to heterogeneous effects of the SPP scholarship by student gender.

Since eligibility for the SPP scholarship is based on a student's Saber 11 test score and Sisben socioeconomic index score, and because there is some noncompliance,[4] we use a multi-score fuzzy regression discontinuity design to estimate the causal impact of SPP receipt on the student outcomes of interest. The empirical strategy we employ allows for the estimation of two types of effects: first, a single pooled effect of marginally qualifying for treatment at any point on the boundary of either score; second, heterogeneous effects from marginally qualifying for the scholarship in the neighbourhood of a given point. For the latter, we estimate effects at three different points: a high test performance point wherein a student marginally qualifies in terms of socioeconomic score, a lower wealth point wherein a student marginally qualifies in terms of test score, and a double-marginal point wherein the student marginally qualifies for the SPP scholarship in terms of just crossing the threshold for both the wealth and test score.

[3]This differs from promise programs that have been introduced in many cities in the United States wherein tuition is covered in full irrespective of graduation; see Bartik, Hershbein, and Lachowska (2021) and Page et al. (2019) for evidence on the effects of two specific promise programs on postsecondary outcomes.

[4]Noncompliance arises since several students that qualify for the scholarship do not elect to take it up, even if they decide to enroll in university. Some students may worry about the financial consequences of the SPP loan if they are unable to graduate. Other eligible students may refuse the program for a variety of reasons including that they have no plans for tertiary education, or prefer vocational education, among other explanations.

To the best of our knowledge, only Cattaneo et al. (2023) has used a multi-score regression discontinuity design to evaluate educational policies. When encountering a situation in which there are multiple numerical threshold requirements to qualify for a policy or program, researchers often assume that they can simply reduce the dimensionality of the problem such that the traditional single-variable regression discontinuity design can be used. For example, in a scenario with two running variables, a researcher can estimate the effect of crossing each threshold separately, ignoring the values of the other running variable in each estimation; alternatively, a researcher can examine the effect of crossing the threshold for one running variable while restricting the sample in a particular fashion using values of the second running variable. An example of the latter in the SPP context would be to estimate the effect of qualifying for the program in terms of test score for those at a given range of lower values of the Sisben score. These approaches are potentially problematic for two reasons: first, they assume a constant treatment effect for each of the qualification thresholds being crossed, irrespective of the value of the second running variable. Second, these approaches reflect endogenous sampling as relevant observations are being excluded from estimation – after all, it is likely that both prior academic performance and socioeconomic status would affect college academic outcomes and that these running variables are not perfectly correlated. By dropping a subset of students who are on the margin of receiving a SPP offer on one of the criteria would exclude those observations from all subsequent analyses.[5] These students are included in multiple-score regression discontinuity designs. In our empirical analysis, we provide statistical evidence of heterogeneity of both treatment effects and how bandwidth parameters along different values of the running variables change, which substantiates the need to take a multi-score regression discontinuity approach when analyzing the effects of the SPP scholarship.

Our paper also contributes to a small but growing literature on the effects of the SPP scholarship. Bernal and Penney (2019) and Laajaj et al. (2022) examine the effect of the SPP program's introduction on the national high school exit exam, finding that the aca-

[5]The observations available for other popular diagnostics associated with the regression discontinuity design such as tests for smoothness of other variables or manipulation at the discontinuity would be affected by this decision to drop observations based on the value of another running variable. These observations also likely change the relevant neighborhood to estimate causal parameters by influencing the optimal bandwidth.

demic performance significantly increased among low-income students. Laajaj et al. (2022) additionally find evidence of increased academic performance of low-income students on the national ninth grade exam, suggesting this program could influence education trajectories. Londoño-Vélez et al. (2020) showed that the introduction of SPP immediately increased the percentage of low-income students that became eligible to attend college, thereby increasing demand for tertiary education. The authors additionally report that colleges increased the supply to accommodate this growth in demand from low-income students that were both eligible and ineligible for SPP.[6] Last, the paper most closely related to ours is Londoño-Vélez et al. (2023), which was conducted independently. The paper differs by using a single-score regression discontinuity design in the empirical analysis, and additionally reports estimates of the effect of SPP receipt on future earnings.[7]

The main findings of this analysis are as follows. We do not find statistically significant results of the scholarship on test scores for the pooled estimates, except for when analyzing the subsample by chosen university: students who attend schools outside of the capital city of Bogota do slightly better on average, while recipients in Bogota do substantially worse. There are consistent positive effects of SPP receipt on Saber Pro test score performance for marginal students at the lower wealth point. Decomposing the Saber Pro score across subject areas, we do not find major differences, with marginal lower wealth students seeing meaningful gains and the other groups seeing limited and statistically imprecise effects. Statistically significant differences are found when examining the impact of the SPP program on the other outcomes under consideration: students are generally much more likely to attend high-cost universities and enroll in high-cost majors, which we define as having tuition of greater than 7,000,000 COP (about 1,750 USD) per year. Marginal lower wealth and high performing students have a significantly higher probability of attending an elite university. Students who qualify for

[6]This result is consistent with Carranza and Ferreyra (2019) finding that higher education institutions in Colombia respond to the economic incentives of policies introduced between 2000 and 2013. Given that higher education expansion involves important trade-offs and research (see e.g., MacLeod and Urquiola, 2015; Urquiola, 2020; Blair and Smetters, 2021) documents limited capacity in elite universities, we are unable to explore whether the SPP program affects the quality of education provided. Such analyses would be necessary to fully evaluate the equity implications of the SPP program. Our results focus on the efficiency of SPP and should be viewed under the light that SPP recipients had a strong economic incentive to have their SPP loan forgiven.

[7]At the time we became aware of this study, we had planned to report that analysis in a companion paper. We elected to maintain that decision in the spirit of a pre-analysis plan to maintain independence.

SPP write the exam earlier and are much more likely to enroll in university and write the Saber Pro exam (as opposed to either not enrolling or enrolling but not yet having written the exit exam). Marginal lower wealth students are more likely to be STEM majors.

Examining the empirical results by subgroups defined by gender, we do find some evidence of important differences. Female students are more likely to attend high-cost universities and majors compared to males, and are also more likely to enroll in top schools. Female students who are SPP recipients also wait less time to write the Saber Pro exam compared to males. For male SPP recipients, there appears to be an increase in probability to enroll in a STEM program, while females do not observe an increase. The propensity to enroll in university and write the Saber Pro exam is similar for males and females, except for marginal high-performance students, where females are much more likely to do so.

This paper contributes to two branches of the economics of education literature. The promise of free tuition conditional on degree completion led to a large change in the number of high-achieving students from lower-income families in Colombia's most elite institutions. This finding mirrors evidence from a field experiment in a single U. S. state that such an early commitment of free tuition more than doubles application and enrollment rates to flagship universities (Dynarski et al., 2021).[8] Burland et al. (2022) present additional results from a second field experiment on a subsequent cohort of students that demonstrate the importance high school students place on financial certainty when making schooling decisions. Last, Montalbán (2023) presents evidence from Spain that stress the benefits for disadvantaged student academic outcomes when financial aid packages also incorporate rigorous academic requirements. Our evidence of changes in both the types of majors and universities attended with the SPP scholarship reinforces that the financial certainty of having costs removed can lead to a widely different sorting of high-achieving, low-income students both within and across institutions. Thus, our finding mirrors evidence using U.S. data presented in Dillon and Smith (2017) that financial constraints play a critical role in explaining college quality choices, particularly for these students. Further, the large positive impacts of SPP are of the same order as evidence from the West Virginia PROMISE merit-based scholarship

[8]Dynarski et al. (2021) interpret their findings as evidence that behavioral biases can explain why these students are initially less likely to attend selective colleges relative to other high achieving students.

(Scott-Clayton, 2011) and Project STAR (Angrist et al., 2009), programs where demanding academic requirements were a component of financial aid. Second, our research contributes to a large literature examining how the effects of college quality on student academic outcomes vary with student ability (see e.g. Light and Strayer, 2000; Black, Daniel and Smith, 2005; Dillon and Smith, 2020; among others). The economic research on the impacts of college quality is surveyed in Lovenheim and Smith (2022) and typically focuses on post-schooling labor market outcomes rather than academic outcomes including college completion (see e.g. Bowen, Chingos and McPherson, 2009; Dillon and Smith, 2020; among others), time to degree (Bound, Lovenheim and Turner, 2010) and academic performance. Prior evidence on college quality on college completion generally focuses on (i) data from institutions where concerns related to student sorting are minimized (e.g., Stange (2012)), or (ii) among studies that use a regression discontinuity design to deal with student sorting focus on administrative data from just a single state (see e.g., Cohodes and Goodman, 2014; Goodman, Hurwitz, and Smith, 2017; Kozakowski, 2020; Bleemer, 2021; among others).[9] Foote and Stange (2022) point out that analyzing administrative data from a single state may lead to biases if students attrit from the sample due to college switching across borders. Our study adds to this literature using nationwide administrative data and contains an academic test score that can be compared across students in different majors both within and across institutions.[10]

This paper is organized as follows. We begin with a discussion of the institutional background of Colombia's higher education sector, the Ser Pilo Paga scholarship, and the data we use in Section 2. In Section 3, we briefly outline the methodology of the multi-score regression discontinuity design and compare it with the textbook single-variable regression discontinuity design approach. Section 4 presents and discusses our empirical results using the multi-score regression discontinuity design, additionally focusing on the broader impacts of the SPP scholarship on institutional admissions. In addition, we present evidence of policy relevant treatment effect heterogeneity and illustrate several of the consequences from

[9]For brevity, we refer the reader to the Lovenheim and Smith (2022) survey for evidence from other research designs including instrumental variables and selection on observables.

[10]We note that national data has been used to evaluate different types of financial aid and grant programs in Chile (Bucarey et al., 2020) and the United States (Eng and Matsudaira, 2021). Note that the pattern of student sorting across institutions in Colombia from SPP differs markedly from what has been found in Chile.

reframing the study by using a single-score RDD. We conclude with a discussion in Section 5.

2. Institutional background and data

Higher education in Colombia is characterized by a strikingly high degree of regional variation and gender difference in employment prospects. Data from the OECD reports that 42% of women in Colombia with below upper secondary attainment were employed in 2021, compared to 72% of those with tertiary attainment. In contrast, the respective figures were 87% and 85% for men. On the former, differences in educational attainment to a large degree reflect differences in education opportunities and are in part due to economic conditions and internal migration patterns. Colombia has a compulsory schooling law with a school leaving age of 16. Primary and secondary education together consist of 11 grades, following which students can elect to attend either a vocational school or tertiary education. Students choose both the college and major when applying. Colombia has a decentralized higher education system with both public and private colleges and universities in operation. Each school has their own cutoff for an admission score on the Saber 11 test that is described in further detail in Subsection 2.2. The cost to attend a higher education institution are set by the institutions themselves, with public institutions being substantially lower than private ones, and the former also offer means-tested tuition fees. Each college and university must meet the minimum quality standards set by the Ministry of Education's Qualified Registry. To differentiate institutions based on their quality, each school has the opportunity to apply for a certificate of High Quality Accreditation, awarded by the National Accreditation Council that has their own criteria and undertakes a peer review evaluation process (see e.g., Camacho et al. (2017) for further details). Roughly 10 percent of all colleges and universities operating in Colombia met the criteria and received High Quality Accreditation on the date that the SPP scholarship program was announced.

Prior to the introduction of the SPP program, researchers have documented substantial segregation in tertiary school enrollment. MacLeod et al. (2017) and Ferreyra et al. (2017) report that high income students are more likely to attend private colleges. Further, they

find that very few high-performing low-income students enrolled in the highly competitive high-quality public schools, with low-income students more likely to sort into low quality public and private schools.

2.1. Saber 11

The Saber 11 is the standardized high school exit exam in Colombia, with the primary purpose being to provide universities and other postsecondary institutions with a standardized measure of academic performance for applicants. Introduced in 1968, this exam became compulsory in 1980. The government agency responsible for designing and administering the test is (El) Insituto Colombiano para el Fomento de la Educación Superior (ICFES), an autonomous entity that works together with the Ministry of Education in Colombia; in addition to administering standardized tests, ICFES's role is to conduct research in measuring the quality of education in the country.

The Saber 11 remains the primary criterion for admission to higher education and is an important determinant in the awarding of scholarships (ICFES, 2010).[11] The Saber 11 exam is conducted at the end of every academic year to students in their last year of high school. However, because there are two academic schedules in the country, there are two (primary) testing dates: a day in the third quarter (normally August) for public school students and private school students who follow the same academic schedule (known as "Calendar A") which runs from February to November, and a day in the first quarter (usually March) for students who attend private schools whose academic calendar follows an international schedule ("Calendar B") which runs from August to June.[12] On each day, the test is administered at the same time across the country. The exam is administered in two separate sessions of four hours and thirty minutes each. The rate of compliance is extremely high, with approximately 97% of students in their last year of education writing the test. While preparing for the test, students also have the option of paying to take a practice version of the Saber 11 test called the PreSaber, which takes place on the usual

[11] Tests that play similar roles in other countries include the Gaokao in China and CSAT in South Korea.

[12] There is a very small group of students who write the test on neither day. These are generally either students with disabilities who require special accommodations to write the exam, or students that are incarcerated or have otherwise been deprived of their liberty.

testing dates.

In 2014 and onwards, the Saber 11 exam is divided into five subject areas: mathematics, language skills, social science, natural science, and English language skills. Each subject area is scored separately. Scaled using Item Response Theory, the individual scores in each section are used to produce an overall score that ranges from 0 to 500, with a mean of 250 and a standard deviation of 50. In addition to questions covering these subjects, test takers are also asked several questions about their demographic characteristics and socioeconomic status.

2.2. Saber Pro

The Saber Pro is the university exit exam in Colombia. The stated purpose of the exam by ICFES is to assess the level of educational quality of higher education institutions in the country. The exam was initially administered in 2003, then known as ECAES,[13] to evaluate learning outcomes for specific majors. It was later expanded to include a common portion and made mandatory in 2010, then being renamed Saber Pro.

The exam is administered once per year on a single testing day.[14] Unlike the Saber 11, which is taken during the last year of high school, the Saber Pro can be taken so long as students have completed 75% or more of the required credits for their degree, and so students can elect to write the exam immediately once they have satisfied the requirement or when they have nearly completed their degree. The exam is divided into two parts: first, a common set of five modules taken by all students; second, either a single or set of modules specific to the choice of major. The five common modules are as follows: quantitative, reading, writing, English, and civics.

The Saber Pro exam is administered in two separate sessions: a morning session and an afternoon session. The morning session covers the common modules, and is four and a half hours long. The afternoon session has the major-specific module(s): if there is a single module, the duration to write the exam is two hours; if there are two or more, the duration to write the exam is four and a half hours. The test scores are scaled using Item Response

[13]ECAES is an acronym for "Exámenes de Calidad de Educación Superior".

[14]The exception was the exam for the year 2020, wherein it was conducted on two separate days, with each student writing their exam on a single day as described above.

Theory, and the overall scale ranges from 0 to 300, with a mean of 150 and a standard deviation of 30.

2.3. Socioeconomic status

While there are two measures of socioeconomic status used in Colombia: strata ("estrato" in Spanish)[15] and Sisben, only the latter influences SPP qualification. The Sisben score is a measure used by the government for the provision of personalized services, such as health and education programs and other subsidies.[16] Specifically, among other things, the purpose of Sisben is to facilitate the classification of potential beneficiaries in a uniform and objective manner, and to support the design and implementation of social programs.

Our analysis uses the Sisben III score for which an initial assessment phase took place between 2009 to 2011, with new households and revision requests occurring afterwards.[17] The score is calculated at the household level, and a government agent visits each household individually to assign a score. The formula for the calculation of the Sisben score is secret but revised periodically. It is determined by variables at the household level that were intended to proxy housing, vulnerability, public services, health, and education. Excluded from the formula for this version of the score are measures of family income and potential income.

The Sisben III score ranges from 0 to 100, with smaller numbers indicating lower socioeconomic status; fractional values are possible. The Sisben score does not expire and does not change if a household moves their primary domicile.[18] Any household can request to be evaluated for a Sisben score, although the percentage of households with Sisben scores falls dramatically at strata 4 and above. Anecdotal evidence suggests that Sisben score revisions were quite rare until the announcement of the SPP program, and that the amount of time to receive a revision is a minimum of six months.

[15]The strata measure is discussed in Bernal and Penney (2019) and is on an integer scale from 1 to 6, with 1 being the lowest and 6 being the highest. Strata is determined entirely by place of residence, so it can be changed by simply moving from one location to another. There are exceptions, however: for example, Indigenous Colombians are considered to be strata 1 regardless of where they reside.

[16]Sisben is an acronym in Spanish for System of Identification of Social Program Beneficiaries.

[17]The process to obtain a Sisben score is explained in more detail in Bernal and Penney (2019).

[18]A new score must be calculated if a new version of the Sisben score is released, such as when the Sisben IV was introduced in 2017. The Sisben IV uses a different formula for its calculation.

2.4. Ser Pilo Paga

On October 1st, 2014, the Colombian government unexpectedly introduced the Ser Pilo Paga program to boost college attainment of low-income, high merit students. The program was quite generous: it included tuition fees with no cap, which was paid directly to the university the student was attending, and a modest stipend for living expenses. The stipend ranged between approximately 1 and 4 times the legal minimum wage in each semester, depending on where the student resided and where the student was enrolled; schools had the option to provide additional support to SPP recipients. The funds provided by SPP to the student were considered a loan, and only upon graduation would this loan be forgiven in which case SPP can be viewed as a traditional scholarship.

To qualify to receive the SPP scholarship, students needed to meet three criteria: first, they needed to reside in a household with a Sisben score at or below a threshold that depended on whether the household resided in a rural area (40.75), an urban area (56.32), or if they lived in one of the country's 14 major cities (57.21).[19] Second, students needed to perform very well on the Saber 11 exam, with the cut-off score being set at approximately the 91st percentile of the overall test score distribution. Since the cut-off score is set only after the test is administered, students do not know the cut-off score in advance. Last, students are only permitted to use the SPP scholarship at a set of universities designated to be "high quality".[20]

Based on the first two criteria, approximately 8% of (Calendar A) students in both 2014 and 2015 in the standard academic schedule were eligible to receive the scholarship; however, some students who qualified for SPP did not necessarily accept it. Those students who received the scholarship were informally referred to as "Pilos". The SPP program provided awards to nearly 40,000 "Pilos" prior to its termination in 2018 that occurred quickly after the 2018 Colombian presidential election.[21] However, this cancellation did not affect the

[19]The list of the 14 major cities in alphabetical order are as follows: Barranquilla, Bogota, Bucaramanga, Cali, Cartagena, Cucuta, Ibague, Manizales, Medellin, Monteria, Pasto, Pereira, Santa Marta, and Villavicencio.

[20]Students needed to apply and get accepted to these schools to receive the SPP scholarship. Attendance was not a requirement and, in our data, we do observe that a very small number of SPP recipients ended up choosing to attend universities that were not on the list of high-quality institutions.

[21]At the time of cancellation, the government claimed this decision was made since the per-student costs

funding to existing recipients.

2.5. Data

We analyze the effect of the Ser Pilo Paga scholarship for the 2014 and 2015 cohorts of high school graduates using a dataset that combines data from ICFES, the DNP, which is a Colombian government agency responsible for public and economic policy, and the MEN, the Colombian Ministry of Education.[22] The dataset was constructed and anonymized for use by the authors.

Summary statistics for our primary regression sample are displayed in Table 1; these are the students that graduated from high school in 2014 or 2015, have Sisben and Saber Pro scores, and enroll in university. The statistics show the full sample and are then divided into two subsamples: those who were awarded the SPP scholarship, and those that were not. We see a slight difference in age at the Saber Pro test between these groups, with SPP recipients being slightly younger. SPP recipients are 6 percentage points more likely to have at least one highly educated parent, which we define as having a professional, technical, or post-secondary degree or certification. Predictably, the SPP group tends to have a lower Sisben score. The elapsed time between writing the Saber Pro and the Saber 11 is also shorter for SPP recipients. There are some notable differences between the SPP and non-SPP groups in terms of choice of university and major: SPP recipients are more likely to attend a high cost university, a top 3 university,[23] and to major in a STEM subject. There are also sizeable differences in the Saber 11 test scores, the Saber Pro test score, and all the Saber Pro subject test scores.

We examine the distribution of the Saber 11 and Sisben score pairs for the observations in the sample on Figure 1. Three symbols on this scatterplot should be noted: a square at the point (0,0), wherein a student just qualifies in terms of both their Sisben score and their Saber 11 test score; a triangle at (5, 0), which we will refer to as the lower wealth point,

to maintain the program were very high, and that funding primarily appeared to flow towards private rather than public universities (Bernal, 2021); see Barrientos and Castañeda (2019) for additional discussion on why the program ended.

[22] In Spanish, the acronym DNP is for "Departamento Nacional de Planeación", and the acronym for MEN is "Ministerio de Educación Nacional".

[23] The country's top 3 universities are Pontificia Universidad Javeriana, Universidad de Los Andes, and Universidad Nacional de Colombia.

wherein students are 5 points below the Sisben qualification threshold but are marginal in terms of test score (note that we invert the Sisben score on this graph for simplicity so that positive numbers indicate the number of points below the threshold a person is); and a bold X at (0, 25), what we refer to as the high test performance point, wherein students are well above (about half a standard deviation) the minimum Saber 11 qualifying score but are marginal in terms of Sisben score. In our empirical analysis, we will report causal parameters at each of these points on the respective thresholds wherein one would qualify for the SPP scholarship.

Figure 1 illustrates the significant amount of support at and around the point (0,0). However, along the Saber 11 dimension, we observe fewer observations past the X symbol: these are students that have scored significantly higher than the required score on the Saber 11 exam. For the Sisben dimension, the number of individuals decreases as we move down the Sisben scores (moving right). Along the cut-offs for the Sisben and Saber 11 dimensions, there are regions with sparse numbers of observations. Last, we note that concerns that this scatterplot is misleading by lacking the visibility of ordered pairs in the sample that constitute duplicate observations (those with the same values of both Sisben score and Saber 11 score) should be limited: in aggregate, only 1.42% of the observations in the sample are such duplicate observations, and the vast majority of these are pairs of identical observations (rather than being triples or more).

Figure 2 displays a hexagonal heat map that shows the density of the test score and socioeconomic score pairs. The lightest hexes each represent approximately 0.335 percent of the sample, while each of the darkest hexes contain approximately 0.012 percent of the sample. The heat map shows the three cross-over points we selected above have sufficient support in their neighbourhoods to calculate meaningful treatment effects at their locations; Cattaneo et al. (2023) warn that there could be excessive extrapolation in the regression discontinuity estimates if areas surrounding the boundary are not sufficiently populated. In terms of the full sample, we see that most of the observations fall about 20 to 60 points below the Saber 11 score threshold, and there is a significant number of points ranging from just above to somewhat below the Sisben socioeconomic score threshold. In addition, there appears to be a small cluster of points at and near the Saber 11 score threshold and ranging

from about 0 to 10 points below the Sisben score threshold.

3. Methodology

In this section, we describe the multi-score regression discontinuity design and how it differs from the single-score regression discontinuity design in practice. To motivate that discussion and how it pertains to our evaluation of SPP, we begin by providing an overview of estimation and inference with the RDD.

3.1. Regression discontinuity design estimation and inference

Sharp regression discontinuity designs correspond to settings where the discontinuity point is not manipulable and compliance to treatment assignment is perfect permitting an unconfoundedness assumption to be plausible. As such, the average causal effect of the treatment at the discontinuity point can be calculated by taking the difference in the conditional expectation of the outcome given the covariate at that point. In the context of the SPP scholarship assignment, this permits us to recover a local intent to treat causal parameter provided by that the standard continuity assumptions summarized in Cattaneo et al. (2020) and Cattaneo and Titiunik (2022) hold.

Policymakers are more likely interested in the effect of using the SPP scholarship, and not the effect of being eligible for the forgivable loan. As such, there is noncompliance since the probability of receiving the SPP treatment does not jump from 0 to 1 at the threshold; this is known as a fuzzy regression design. In an important paper, Hahn et al. (1999) use the relationship between sharp and fuzzy designs to draw a link with the encouragement design allowing them to exploit the connection to the instrumental variables estimator showing that, under slight variants to the assumptions in Angrist et al. (1996), one can recover a local average treatment effect. In our setting, this is the average effect of SPP receipt on the outcome of interest only for compliers at the discontinuity point. Formally, if a cutoff point is given by c, the causal parameter is given by

$$\tau_1 = \frac{lim_{x \downarrow c} E[Y|X = x] - lim_{x \uparrow c} E[Y|X = x]}{lim_{x \downarrow c} E[SPP|X = x] - lim_{x \uparrow c} E[SPP|X = x]} \tag{1}$$

where Y reflects the academic outcome and X is the running variable. However in our context, eligibility for SPP is determined by two variables, the Sisben and Saber 11 scores, and there is no longer a single unique point where the probability of being eligible for treatment jumps from 0 to 1. The above setting has been coined in Cattaneo et al. (2020) to be a multi-score regression discontinuity design and is an extension of the above discussed standard regression discontinuity design, which we refer to here as single-score regression discontinuity design.

Without loss of generality, consider the case as we do in this paper wherein there are two running variables, each with its own threshold, and both thresholds must be met or exceeded in order to qualify for treatment: label the running variables as S_1 and S_2, the values of the running variables s_1 and s_2, and their respective thresholds as ${s_1}^*$ and ${s_2}^*$.[24] Then, treatment occurs at two different boundaries: first, the boundary s_1 such that $s_2 \geq {s_2}^*$, second, the boundary ${s_2}^*$ such that $s_1 \geq {s_1}^*$. If either of the running variables are continuous, then there are an infinite number of crossover points into treatment.[25] By having more than one running variable, there are several implications for estimation and inference that are reviewed in Section 5.2. of Cattaneo et al. (2023). Further, since there now exist multiple different crossover points where one gains eligibility to SPP generating a boundary plane, a causal effect of crossing the boundary can be calculated at any point along that boundary.

Similarly, a suite of local average treatment effects for each pair s_1 and s_2 can be estimated with a fuzzy multi-score regression discontinuity design. To identify this causal parameter for compliers, Choi and Lee (2023) discuss the need for (semi-) strong monotonicity on the single instrument for the SPP program.[26] While a suite of local average treatment effects can be estimated, we do not adjust statistical inference for multiple testing since under the fuzzy RD design, over-rejected tests and under-covered confidence intervals can arise if either the sample size or the proportion of compliers are small for some subsamples at some s_1 and s_2 pairs. That is, the strength of the first stage could vary across s_1 and s_2 pairs, leading the estimators to have a non-normal distribution and consideration of alternative methods

[24]For simplicity of exposition, we reverse-code the Sisben score in our analysis such that higher scores denote poorer individuals. This has otherwise no effect on the empirical analysis.

[25]If the scores are both discrete and subject to a grid (e.g. a grid size of 0.1), there is still potentially a very large number of crossover points.

[26]Note that the restriction of a single instrument is important, and both Mogstad et al. (2021) and Heckman and Pinto (2018) consider what can be identified with an unordered treatment when a researcher has access to multiple instruments.

for weak inference (see e.g. Moreira (2003) or the recent survey by Andrews, Stock and Sun (2019)).

In both sharp and fuzzy multi-score regression discontinuity designs, researchers can report causal effects that fall into one of two categories: first, a pooled effect that is the effect of crossing the boundary plane at any point; second, a location-specific effect of crossing the boundary in the neighbourhood of that specific point, which we refer to in this paper as a local average treatment effect at points s_1 and s_2. To calculate either causal parameter requires one to first measure the distance between an individual observation in the sample and either the closest boundary in the case of the former, or to the specific point on the boundary plane in the case of the latter. For the pooled effect, the distance metric S is calculated as follows: if for the point (s_1, s_2) we have $s_1 \geq {s_1}^*$ but $s_2 \leq {s_2}^*$, the distance is calculated as $s_2 - {s_2}^*$; if $s_2 \geq {s_2}^*$ but $s_1 \leq {s_1}^*$, the distance is $s_1 - {s_1}^*$; if $s_1 \geq {s_1}^*$ and $s_2 \geq {s_2}^*$, the minimum of $(s_1 - {s_1}^*)$ and $(s_2 - {s_2}^*)$; if $s_1 \leq {s_1}^*$ and $s_2 \leq {s_2}^*$, the Euclidian distance metric $((s_1 - {s_1}^*)^2 + (s_2 - {s_2}^*)^2)^{1/2}$ is used. For the local average treatment effect at a given point $({s_1}', {s_2}')$ on the boundary plane, all distances are calculated using the Euclidian distance metric. For both categories of treatment effects, untreated observations have negative distances, and treated observations have positive distances. Note that the calculated distance metrics for the observations in the sample are different for each crossover point of interest and for the pooled effect, and thus a separate regression discontinuity design and distance metrics must be estimated for each.

3.2. Comparison with the single-score approach in a multi-score environment

In practice, researchers (see e.g., Jacob and Lefgren (2004), Matsudaira (2008), Papay et al. (2010), among others) often recast a multi-score RDD as a series of single-score RDD in both sharp and fuzzy settings.[27] To keep the discussion brief, we focus on the sharp setting wherein access to a treatment requires crossing thresholds on two running variables as in the case of the multi-score RDD we consider in this paper. Rather than follow the

[27] Another option is to assign units based on a new running variable that is defined as the running variable that is closest to its respective cutoffs see e.g. Gill et al. (2007), Battistin et al. (2009), or Clark and Martorell (2014), among others. This approach of shortest distance would correspondingly exclude all observations that are assigned to treatment via the second assignment variable and cutoff.

approach introduced by Cattaneo et al. (2020), researchers use a choice based subsample of the original data so that only a single running variable determines treatment eligibility, and then estimate a single score regression discontinuity design for the effect at the threshold of the remaining running variable. Thus, one regression discontinuity design is estimated for the cutoff ${s_1}^*$ using a subsample of the data such that $s_2 \geq {s_2}^*$, and similarly another RDD is estimated for the causal effect at the cutoff ${s_2}^*$ using a subsample of the data such that $s_1 \geq {s_1}^*$.

When undertaking the above analyses, researchers implicitly assume that the two running variables are independent of each other. Further, researchers report a single causal parameter and do not allow the estimated causal parameter to systematically vary over the distribution of the remaining running variable. For example, perhaps crossing ${s_1}^*$ for small values of s_2 produces very large effects, but the effects of crossing ${s_1}^*$ for large values of s_2 are near zero. These kinds of heterogeneities are masked under the single-score approach, and such differences in outcomes can have significant policy implications

The above research strategy faces the concern that either or both the $s_1 \geq {s_1}^*$ and $s_2 \geq {s_2}^*$ subsamples are choice-based and should not be treated as random samples; this influences the internal validity of the estimated causal parameter. In addition, using subsamples defined on crossing a threshold on the other running variables would have consequences for empirical practice ranging from bandwidth selection to specification tests related to manipulation, where the consequences increase the more the running variables are correlated with each other. To illustrate, consider the calculation of crossing the threshold at the cutoff ${s_1}^*$ where one removes all observations from the original data where $s_2 < {s_2}^*$, even if such observations are close to the threshold ${s_1}^*$. The conditional expectations used to calculate the causal parameter presented in equation (1) would now be represented as

$$\tau_1 = \frac{lim_{s_1 \downarrow {s_1}^*} E[Y|S_1 = s_1, s_2 \geq {s_2}^*] - lim_{s_1 \uparrow {s_1}^*} E[Y|S_1 = s_1, s_2 \geq {s_2}^*]}{lim_{s_1 \downarrow {s_1}^*} E[SPP|S_1 = s_1, s_2 \geq {s_2}^*] - lim_{s_1 \uparrow {s_1}^*} E[SPP|S_1 = s_1, s_2 \geq {s_2}^*]} \tag{2}$$

that likely differs if there are heterogeneous treatment effects from the original equation (1)

$$\tau_1 = \frac{lim_{s_1 \downarrow {s_1}^*} E[Y|S_1 = s_1] - lim_{s_1 \uparrow {s_1}^*} E[Y|S_1 = s_1]}{lim_{s_1 \downarrow {s_1}^*} E[SPP|S_1 = s_1] - lim_{s_1 \uparrow {s_1}^*} E[SPP|S_1 = s_1]} \tag{3}$$

Researchers would need to implicitly assume that observations are dropped randomly at the cutoff point to have the same interpretation. Otherwise, bias is introduced if the effect on outcomes changes discontinuously at the second running variable sample restriction, if one were to recast to a single-score RDD. Researchers assume this sort of effect is absent if they undertake such a conversion of a multiscore RDD and intuitively, we would argue that bias is much more likely to exist when sampling on the Saber 11 frontier relative to the Sisben frontier. Yet, even in the absence of such an effect causing bias, there is a cost to efficiency since the number of observations used to calculate the conditional expectations at the cutoff point in equation (2) would now be lower than equation (1'). Further, by using a smaller data set may lead to the choice of a larger bandwidth to minimize the mean squared error.[28] In other words, one is only estimating a causal parameter at a specific discontinuity frontier $s_1 \geq {s_1}^*$. Last, to estimate a combined effect from two single-score RDD would also require a continuity in expectations of both potential outcomes ($d = 0, 1$) that can be expressed as:

$$lim_{s_1 \uparrow {s_1}^*} E[Y(d)|s_1 = {s_1}^*, s_2 \geq {s_2}^*] = lim_{s_1 \downarrow {s_1}^*} E[Y(d)|s_1 = {s_1}^*, s_2 \geq {s_2}^*]$$

$$lim_{s_2 \uparrow {s_2}^*} E[Y(d)|s_1 \geq {s_1}^*, s_2 = {s_2}^*] = lim_{s_2 \downarrow {s_2}^*} E[Y(d)|s_1 \geq {s_1}^*, s_2 = {s_2}^*]$$

These assumptions could be avoided by undertaking the multiscore RDD approach. Last, if the running variables are correlated, then the impact of considering the truncation on one running variable for bandwidth selection on the remaining variable becomes more important the greater is the correlation, with boundary effects becoming an increasing concern. In our data, the raw Pearson correlations between the Sisben and Saber 11 scores is -0.1030 and -0.0190 for the 2014 and 2015 graduating cohorts, respectively.

In applications of regression discontinuity design, researchers often present results from a series of specification tests to strengthen the credibility of their estimate.[29] First, they

[28]This result is well known in the regression discontinuity literature that the size of the bandwidth is much larger when the cutoff point is located in the tails of a bell-shaped distribution relative to the case where this point is near the center of the distribution. We also note that more bandwidth is needed to be selected for a fuzzy RDD relative to a sharp RDD.

[29]There are additional specification tests discussed in Cattaneo and Titiunik (2022) or Lee and Lemieux (2010) that may be relevant. This also includes Lee and Card (2008) who study the case wherein a running variable is discrete such as with the Sisben score and its consequences for statistical inference. We follow their guidance related to adjusting the standard errors to provide more conservative inference.

attempt to convince readers how that no other explanatory variable with the exception of the treatment of interest changes discontinuously at the cutoff point of the running variable. This test becomes less convincing the more a tested covariate is correlated with the running variable that was truncated in order to convert the research design to a single-score RDD. Second, researchers conduct specification tests to convince readers that it is unlikely that there is manipulation of the running variable under consideration. A common test is to show continuity of the density of the running variable at the discontinuity point, against the alternative of a jump in the density function at that point. Yet, there are no guarantees that the observations dropped to create this subsample would exhibit the same patterns, and the test needs to account for the endogenous sample construction. In our empirical results, we show evidence that cast doubt on reframing the multi-score RDD as a series of single-score RDD regressions in the context of our application.

4. Empirical Analysis

We begin the analysis by analyzing the effects of the SPP scholarship on Saber Pro scores; the results are displayed in Table 2. As previously stated, we consider a pooled effect, and local average treatment effects at the double-marginal point (0, 0), the lower wealth point (5, 0), and the high test performance point (0, 25). In the overall sample, we do not find any statistically significant effects except at the lower wealth point, which displays an increase of 0.13 standard deviations. This effect size is slightly smaller than the estimates seen in the small class size literature; thus, we categorize the effect size as moderate.

We note the heterogeneity of the estimated effects at the different crossover points: for example, the (25, 0) coefficient of -0.101 is not in the 95% confidence interval for the (0, 0) effect or the (0, 5) effect; moreover, the (0, 0) coefficient of 0.042 is not in the 95% confidence interval for the (0, 5) effect. Large differences in bandwidths are also observed for this regression, with the (0, 5) point having an optimal bandwidth of 21.25, while the (25, 0) point has a much larger bandwidth of 35.2.[30] These results confirm that the multi-score

[30]It is important to note that similar bandwidths in a multi-score regression discontinuity design for different crossover points do not imply that the same observations are being included in the estimations of their causal effects. This is because each crossover point has a different set of distance metrics calculated for the observations in a sample: an observation could be within the bandwidth of one crossover point but not

regression discontinuity approach is useful in the context of calculating causal effects of the SPP scholarship program.

We examine four subgroups defined on predetermined variables: male and female students, and students that resided in either urban or rural areas while attending high school. For males and females, we see a similar increase for the lower wealth students as we see in the full sample, but we additionally observe a significant effect for males at the high test performance point. Students that resided in rural areas do not appear to benefit from the program in terms of Saber Pro scores.

We also conduct an analysis examining subgroups on variables determined by the student choice after they are informed of whether they receive the scholarship: whether they attend a private or public school, whether they attend an elite (top 3) university, and whether the university they enroll in is located in Bogota. It is important to note that the estimated effects for these are affected by selection and therefore should be interpreted as average observed differences for these subgroups rather than causal effects since students make their choices for membership in these groups in the regression after treatment. We observe higher test scores at the low wealth and high test performance points in private schools, and no effects for public school attendees of SPP scholarships are found. Higher test scores for SPP recipients are found at the lower wealth point for non-top 3 schools, while there are no observable differences in scores for elite (top 3) universities. Scores tend to be higher at the double-marginal point and the lower wealth point for recipients attending universities outside of Bogota. For schools in Bogota, the test scores for SPP scholarship recipients are much lower on average, although they are higher at the high test performance point.[31]

Table 3 presents estimates by subject area for the full sample. The only consistent result is that lower wealth students tend to see mild to moderate increases in test score performance in all subjects except civics, with the only other results being that double-marginal students see benefits in English scores, and a very small but precisely estimated effect is seen in

the other, even though both crossover points have the same bandwidth for the calculation of their causal effects.

[31]This may have been the result of less academically prepared students disproportionately moving to Bogota; other factors may include having to deal with the additional challenges of lack of family support or adjusting to life in a larger city. Another possibility is a big-fish-in-little-pond effect wherein SPP students attending school outside of Bogota would perform stronger relative to their peers as a result.

quantitative scores for high test performance students.

Given the above results, we conjecture that the financial resource constraint is important in test score performance because the lower wealth students seem to be the only group that is consistently seeing increases in test scores. Perhaps these increases observed for lower wealth students originate from either reducing their workload or no longer working in order to help pay for their post secondary studies. The general lack of results for the high test performance point make intuitive sense: these students are already very high achievers, and are thus likely already making sacrifices in other areas to maximize their academic achievement and therefore are likely not changing their studying behaviour as a result of receiving the grant. The results for the double-marginal students can almost be argued to be a precisely estimated zero and seem to indicate that the grant is likely not significantly changing their studying behaviour.

The effect of receipt of the scholarship on other university outcomes is displayed on Table 4. The outcomes we examine are whether a student attends a high cost university and program, which we define as having tuition of greater than 7,000,000 COP per year, whether a student chooses to enroll in a top 3 school, the delay between writing the Saber 11 exam and the Saber Pro exam in calendar years ("When write exam"), whether the student majors in STEM, whether the student enrolls in the same program that they initially intended to at the time of writing the Saber 11 exam ("Expected program", which is equal to 1 if their expectations match their outcome),[32] and one more variable which we call "Write exam".

This final outcome variable is a binary indicator that is set equal to 1 if a student has a Saber Pro test score in the data, and equal to 0 otherwise. Thus, if a student has a value of 1 in this variable, they have enrolled in university, have obtained at least 75% of the necessary credits to graduate, and chose to write the test. However, if the value of this variable is equal to 0, one of three things may have occurred: first, the student is currently enrolled in university but either cannot or has not yet chosen to write the test (since they may have the necessary credits to do so but have simply chosen not to); second, the student enrolled in university and later dropped out; third, the student has chosen not to enroll in university

[32]Only a random subsample of students in the 2014 cohort were asked this question at the time of writing the Saber 11, and thus the sample size for this estimate is smaller.

at all. Some may prefer to think of this outcome variable as indicating graduation, as it is reasonable to conjecture that it is very unlikely that a student writes the Saber Pro exam but does not later finish their degree.

Receipt of the scholarship appears to have a very large effect on the decision to enroll in high cost schools, with the probability increasing by 31 percentage points for the pooled estimate, with similar increases at the double-marginal point and lower wealth point; students at the high test score performance point appear to have an even larger increase. The probability to attend a top 3 school does not rise in general when receiving the SPP scholarship, but the lower wealth point sees a small increase of 1.7 percentage points, and the high test score performance point sees a much larger increase at 6.9 percentage points. The delay between writing the Saber 11 and Saber Pro exam shrinks by approximately 0.2 years, except for the high test performance point, which observes a decrease of about 0.28 years. Only those at the lower wealth point see an increase in probability to enroll in STEM programs, and receipt of the SPP scholarship does not appear to be related to the probability of enrolling in the program one thought they would in high school. The probability of writing the Saber Pro exam increases substantially, about 21 percentage points on average, with a larger effect at the lower wealth point and a smaller effect for the high test performance point.

Impacts of the SPP scholarship by student gender are presented in Table 5. The effect on the probability of attending a high cost program is higher for female recipients of the SPP scholarship. For the effect on elite university attendance, females observe a small average effect and an effect at the lower wealth point of about 2.6 percentage points, while males see a large effect only at the high test performance point. The effect on how quickly a student writes an exam is larger for female than male students and precisely estimated for all three points, and is only imprecisely estimated for males at the lower wealth point. There is an average increase in STEM enrollment of 5.4 percentage points for males, though we do not observe precisely estimated increases at the three cross-over points of interest. Effects are not observed for female students. In terms of whether the expected major matches the realized choice in university, we see no effects except for females at the high test performance point, who observe a 52.5 percentage point decrease; this signifies a very large change in the major they expect to enroll in and what major they are registered in at the time they write

the Saber Pro exam. Finally, the probability of enrolling in university, accumulating the necessary credits, and then writing the Saber Pro exam is similar for males and females, but the effect for high test performance males is only about half that of females.

4.1. Robustness and validity

Single-score regression discontinuity designs assume that the causal effect of crossing a given threshold of a running variable on an outcome of interest is due to the change in treatment status that occurs as a result, and that any changes in the outcome are not due to differences in predetermined variables on either side of the threshold; thus, if the distribution of one or more predetermined variables changes at the threshold, it can bring the causal interpretation of the estimate into question. This sort of logic still holds true in the case of multi-score regression discontinuity, although the situation is not completely analogous because there is a large if not infinite number of cross-over points in this scenario. Because of this, a number of these points will likely reject the null hypothesis of balanced covariates in robustness checks simply by chance. These rejections may be especially more likely at points on the treatment border wherein there is a low amount of support, as such locations will contain substantially more noise. Thus, we argue that what is relevant is to examine covariate balance at the cross-over points of interest, which in our case are the three aforementioned points: the double marginal point, the low wealth point, and high test score performance point. Performing this exercise, we see no statistically significant differences in gender, age, rural household location, and parental education status. We also examine covariate balance along the entire border of the discontinuity and fail to find any statistically significant differences in these variables.

Another standard robustness check is to examine the density of the running variables because a discontinuity at the threshold would suggest possible manipulation, especially in situations where treatment is seen by the agent as desirable (or undesirable). There are no differences with this check when using a multi-score regression discontinuity design, other than the fact that each dimension of the running variables must be verified for continuity. However, we express significant doubt that manipulation is possible for either of our running variables of interest in our analysis. For the Saber 11 score, the score to qualify for the

SPP scholarship is not known in advance of taking the test; it is in fact dependent on the distribution of scores after it takes place, and so students are incentivized to perform as well as possible to qualify for the SPP scholarship, absent any other incentives. Given the method in which the cut-off is chosen, it is ipso facto not possible for students to manipulate their scores, even if they possessed the ability to do so to a great extent. For the Sisben score, that the formula for its calculation is secret and with the granularity of the score (the score varies up to two decimal places), fine manipulation of the score around the threshold in this circumstance also seems extremely unlikely if not impossible.

Nevertheless, we conduct these robustness checks for continuity as a matter of course. Since any potential manipulation of the Sisben score would occur around the specific thresholds according to residence, and similarly, manipulation of the test score at the threshold would occur within the year in which the test was taken (since the qualifying score changes from year to year), we examine all of these scores individually. We find no statistical evidence of possible manipulation of these variables at any conventional level of significance; the figures for the manipulation tests are contained in the Appendix.

One concern about the effect of the SPP scholarship was that some social programs in Colombia have similar or identical Sisben score thresholds for qualification (Laajaj et al., 2022), and so it is possible that other programs could be at least partly driving the effects on outcomes. This has been addressed in Bernal and Penney (2019), who examine whether the introduction of the SPP scholarship led to an increase in Saber 11 test scores in 2015 using a regression discontinuity design and found a clear increase in performance at the Sisben thresholds. A robustness check using 2014 Saber 11 test score data, which should not be affected since the SPP program was announced only after the test was written, found no effect on test scores at the qualification thresholds; thus, we conclude that it is very unlikely that these other programs could be having effects on educational outcomes. In the context of this paper, an additional safeguard is the fact that the multi-score regression discontinuity randomly assigns students on both sides of the Saber 11 test score threshold, thus balancing recipients of these other programs equally between the treated and untreated groups along this dimension.

5. Conclusion and discussion

We interpret our evidence in this paper as suggesting that financial constraints are important in explaining the behavior of high-achieving, low-income students in Colombia. However, we note that the results of the analysis are not only due to the relaxing of financial constraints, but also the prospect of a severe financial penalty being imposed for failing to graduate since the SPP program was a conditional loan that is only forgiven upon graduation. The rate of take-up of the SPP program being below 100%, even for those who later went on to enroll in university, suggests that this was both an important and salient feature of the program.

The results of this paper show that, for SPP recipients relative to non-recipients, there is an observed increase in attendance at high cost universities, as well as an increase in enrollment and persistence (which we proxy by students taking the Saber Pro exam): this suggests that there exist significant financial constraints for low income, high achieving students.[33] Assuming that the Saber 11 is a noisy measure of ability, financial constraints for these students therefore result in the potential misallocation of human capital in the broader economy, as they are either enrolling in lower quality universities or not continuing their studies at all. While students have access to student loan programs and other financial resources should they fail to qualify for SPP, the prospect of obtaining a large amount of student debt is likely a significant obstacle to deciding to enroll in a high cost university. Tuition at the many highly ranked private universities can be quite substantial: for example, at Pontificia Universidad Javeriana and Universidad de los Andes, tuition can reach almost 25,000,000 COP per year as a domestic undergraduate student studying economics. Another potential source of resource misallocation that was found in the analysis is the fact that SPP recipients write the Saber Pro exam approximately two to four months earlier than non-recipients: if this translates to students graduating more quickly, this implies that students are entering the labour market more slowly as result of financial constraints, and thus there is a minor loss to overall economic output as a result.

An adjacent concern is that SPP scholarship winners had no incentive to be mindful

[33] We again note the very large increase in the probability of enrolling in one of the three elite Colombian higher education institutions that students at the high test performance point observe.

of tuition costs (so long as they intended on graduating) because the award covered the entire cost of tuition. Thus, universities preferred by winners had strong incentives to raise tuition, especially if they had market power: there exists anecdotal evidence that some private universities raised their tuition fees at least partly motivated by the lack of a cap on the tuition award for Ser Pilo Paga awardees. This suggests that future scholarship programs may want to consider capping tuition fees at a certain level, or perhaps making the scholarships available only for students enrolled at public institutions.

The scholarship had limited effects on student test scores, except for students at the lower wealth point. Thus, student funding is likely not a substantial obstacle for academic performance for most students, and policymakers shouldn't expect student performance to substantially improve when providing students with additional financial resources, except for perhaps those of generally lower socioeconomic status, who may be working during their degree to the detriment of their academic performance.

Breaking down the results on the outcomes by gender, in general, the effects were found to be larger for females compared to males. We speculate that this is due to the generally higher level of risk aversion in females (Borghans et al., 2009; Cortes et al. 2023; among others); hence, the results of this paper show that females are being more negatively impacted by financial constraints compared to males. One important difference found in the analysis is that STEM enrollment increases for male but not female SPP awardees: this suggests that policies designed to increase female STEM enrollment in university should focus on making aid designed for this purpose conditional on the choice of major.

Attracting more high-achieving, low-income students to attend high quality universities is often viewed as providing them a pathway to improve their social mobility. Perhaps the most ambitious program that has been developed worldwide with this goal in mind is Colombia's Ser Pilo Paga and we examine its impact on several academic outcomes. We also find evidence that suggests, in its absence, students were unable to secure the resources to finance their preferred college education. This finding differs from evidence in the United States where these students are shown to face information barriers and are more likely to enroll in colleges and degree programs that have lower returns. Not only did SPP change student behavior, but immediately Colombia's colleges and universities responded

by admitting more students. Whether this influenced student-faculty interactions, teaching quality, and the student experience and peer interactions is a topic for further study.

Our empirical evidence of the impacts of SPP on academic outcomes is drawn from a research design that considers how the scholarship is offered to eligible students in practice. Using a fuzzy multi-score regression discontinuity design does rely on a stronger assumption of monotonicity to recover a suite of causal parameters relative to the traditional single score fuzzy regression discontinuity design. That said, the multi-score design allows the researcher to use the full data and does not require the need for an endogenous sampling restriction. Our analysis illustrates that there are serious consequences for policy evaluation from re-casting SPP as if it offered to eligible participants only on the basis of a single score. Yet, methodological work is required to extend popular regression discontinuity diagnostic tests to the multiple score setting as well as consider inference on treatment effect heterogeneity. This as well as examining the impacts of the SPP program on labor market outcomes present an agenda for future research.

Table 1: Summary statistics

	Full		No SPP		SPP	
	Mean	S.d.	Mean	S.d.	Mean	S.d.
Female	0.640	0.480	0.666	0.472	0.494	0.500
Age	22.007	1.065	22.052	1.070	21.750	0.997
Rural	0.123	0.328	0.125	0.331	0.108	0.310
Highly educated parent	0.359	0.480	0.350	0.477	0.410	0.492
Sisben	41.697	19.219	43.094	19.624	33.739	14.333
Received SPP	0.149	0.356	0	0	1	0
Saber Pro delay time	4.985	0.693	5.019	0.703	4.788	0.600
High cost	0.106	0.307	0.044	0.205	0.456	0.498
Top 3 university	0.027	0.161	0.012	0.109	0.110	0.313
Majored in STEM	0.264	0.441	0.230	0.421	0.462	0.499
Saber 11 global score	0	1	-0.234	0.870	1.331	0.554
Saber Pro global score	0	1	-0.189	0.920	1.075	0.719
Saber Pro quantitative score	0	1	-0.167	0.933	0.950	0.825
Saber Pro reading score	0	1	-0.159	0.949	0.904	0.780
Saber Pro writing score	0	1	-0.070	0.979	0.397	1.027
Saber Pro English score	0	1	-0.155	0.943	0.881	0.848
Saber Pro civics score	0	1	-0.150	0.951	0.853	0.833
Number of observations	89,936		76,511		13,425	

Notes: Summary statistics calculated from sample data. S.d. denotes standard deviation. See Section 4.1 for more details.

Table 2: Test score regressions, Saber Pro global score

	Global	(0,0)	(5,0)	(0,25)
Full sample	-0.003	0.042	0.130**	-0.101
	(0.026)	(0.043)	(0.044)	(0.040)
Male	-0.004	0.029	0.136*	0.126*
	(0.041)	(0.058)	(0.070)	(0.089)
Female	0.004	0.029	0.125**	-0.019
	(0.028)	(0.044)	(0.045)	(0.068)
Urban	0.004	0.051	0.169**	0.046
	(0.030)	(0.036)	(0.061)	(0.064)
Rural	-0.121	-0.287	-0.044	0.224
	(0.096)	(0.165)	(0.130)	(0.675)
Private	0.006	0.048	0.141*	0.081**
	(0.032)	(0.044)	(0.068)	(0.067)
Public	-0.120	-0.212	-0.080	-0.999
	(0.078)	(0.150)	(0.174)	(0.103)
Non-top 3 school	-0.001	0.053	0.132**	0.023
	(0.026)	(0.048)	(0.051)	(0.062)
Top 3 school	-0.187	-0.242	-0.210	0.185
	(0.116)	(0.207)	(0.303)	(0.153)
Not in Bogota	0.084	0.164**	0.222**	0.000
	(0.032)	(0.060)	(0.050)	(0.086)
Bogota	-0.173**	-0.141	-0.035	0.140**
	(0.031)	(0.050)	(0.057)	(0.089)

Notes: * denotes statistical significance at the 5% level and ** for the 1% level. Inference is conducted using robust bias-corrected statistics; see Section 4.2 for more details.

Table 3: Test score regressions, Saber Pro subject scores

	Global	(0,0)	(5,0)	(0,25)
Quantitative	-0.056	-0.001	0.056*	-0.007*
	(0.030)	(0.049)	(0.041)	(0.056)
Reading	-0.008	0.025	0.075*	-0.138
	(0.025)	(0.052)	(0.045)	(0.049)
Writing	0.044	0.058	0.062*	0.126
	(0.035)	(0.048)	(0.038)	(0.084)
English	0.022	0.110*	0.179*	0.173
	(0.033)	(0.054)	(0.078)	(0.111)
Civics	-0.038	-0.045	0.037	-0.042
	(0.030)	(0.050)	(0.047)	(0.059)

Notes: * denotes statistical significance at the 5% level and ** for the 1% level. Inference is conducted using robust bias-corrected statistics; see Section 4.2 for more details.

Table 4: Test score regressions, other university outcomes

	Global	(0,0)	(5,0)	(0,25)
High cost	0.310**	0.295**	0.311**	0.389**
	(0.013)	(0.017)	(0.018)	(0.035)
Top 3 school	0.016	0.018	0.017**	0.069**
	(0.009)	(0.012)	(0.007)	(0.025)
When write exam	-0.224**	-0.220**	-0.198**	-0.280**
	(0.020)	(0.037)	(0.039)	(0.050)
Major in STEM	0.009	0.024	0.049*	-0.051
	(0.016)	(0.024)	(0.024)	(0.034)
Expected program	0.027	0.018	0.123	-0.155
	(0.083)	(0.076)	(0.127)	(0.185)
Write exam	0.206**	0.274**	0.302**	0.142**
	(0.011)	(0.026)	(0.024)	(0.021)

Notes: * denotes statistical significance at the 5% level and ** for the 1% level. Inference is conducted using robust bias-corrected statistics; see Section 4.2 for more details.

Table 5: Test score regressions, other university outcomes, by gender

	Global	(0,0)	(5,0)	(0,25)
High cost				
Male students	0.290**	0.247**	0.279**	0.367**
	(0.021)	(0.031)	(0.033)	(0.054)
Female students	0.321**	0.355**	0.341**	0.406**
	(0.016)	(0.034)	(0.024)	(0.064)
Top 3 school				
Male students	0.012	-0.003	0.022	0.079*
	(0.015)	(0.012)	(0.019)	(0.040)
Female students	0.024**	0.026	0.028**	0.067
	(0.009)	(0.016)	(0.010)	(0.038)
When write exam				
Male students	-0.189**	-0.197**	-0.103	-0.290**
	(0.034)	(0.046)	(0.071)	(0.061)
Female students	-0.239**	-0.237**	-0.251**	-0.290**
	(0.025)	(0.052)	(0.047)	(0.066)
Major in STEM				
Male students	0.054*	0.075	0.081	-0.021
	(0.030)	(0.040)	(0.042)	(0.054)
Female students	-0.004	-0.021	0.008	-0.015
	(0.018)	(0.037)	(0.028)	(0.061)
Expected program				
Male students	0.130	0.111	0.504	0.130
	(0.113)	(0.213)	(0.277)	(0.243)
Female students	-0.108	0.032	0.061	-0.525
	(0.124)	(0.299)	(0.157)	(0.293)
Write exam				
Male students	0.191**	0.253**	0.313**	0.117**
	(0.015)	(0.025)	(0.036)	(0.029)
Female students	0.215**	0.256**	0.283**	0.245**
	(0.019)	(0.031)	(0.031)	(0.047)

Notes: * denotes statistical significance at the 5% level and ** for the 1% level. Inference is conducted using robust bias-corrected statistics; see Section 4.2 for more details.

References

[1] Borah, Bijan J., and Anirban Basu. (2013) "Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence," *Health Economics*, vol. 22: 1052-1070.

[2] Andrews, Isaiah, James H. Stock, and Liyang Sun. "Weak instruments in instrumental variables regression: Theory and practice." *Annual Review of Economics* 11 (2019): 727-753.

[3] Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association* 91.434 (1996): 444-455.

[4] Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. "Incentives and services for college achievement: Evidence from a randomized trial." *American Economic Journal: Applied Economics* 1.1 (2009): 136-163.

[5] Barrientos, Julia Alegre, and Camilo Peña Castañeda. El fin de Ser Pilo Paga: ¿y ahora qué?. El Tiempo. Retrieved from https://www.eltiempo.com/vida/educacion/lo-que-podria-ocurrir-despues-del-fin-de-ser-pilo-paga-265972 (accessed March 31st, 2019).

[6] Battistin, Erich, et al. "The retirement consumption puzzle: evidence from a regression discontinuity approach." *American Economic Review* 99.5 (2009): 2209-2226.

[7] Bernal N., Gloria L. (2021). On improving education opportunities: Preferences and performance of High School students in response to scholarships, information, and co-education. Chapter 4. [Doctoral Thesis, Maastricht University]. Boekenplan. https://doi.org/10.26481/dis.20210623gn

[8] Bernal, Gloria L., and Jeffrey Penney. "Scholarships and student effort: Evidence from Colombia's Ser Pilo Paga program." *Economics of Education Review* 72 (2019): 121-130.

[9] Black, Dan, Jeffrey Smith, and Kermit Daniel. "College quality and wages in the United States." *German Economic Review* 6.3 (2005): 415-443.

[10] Balir, Peter Q., and Kent Smetters. Why Don't Elite Colleges Expand Supply?. No. w29309. National Bureau of Economic Research, 2021.

[11] Bleemer, Zachary I. On the Meritocratic Allocation of Higher Education. University of California, Berkeley, 2021.

[12] Borghans, Lex, et al. "Gender differences in risk aversion and ambiguity aversion." *Journal of the European Economic Association* 7.2-3 (2009): 649-658.

[13] Bound, John, Michael F. Lovenheim, and Sarah Turner. "Why have college completion rates declined? An analysis of changing student preparation and collegiate resources." *American Economic Journal: Applied Economics* 2.3 (2010): 129-15

[14] Bowen, William G., Matthew M. Chingos, and Michael S. McPherson. Crossing the finish line: Completing college at America's public universities. Princeton University Press, 2009.

[15] Bucarey, Alonso, Dante Contreras, and Pablo Muñoz. "Labor market returns to student loans for university: Evidence from Chile." *Journal of Labor Economics* 38.4 (2020): 959-1007.

[16] Burland, Elizabeth, Susan Dynarski, Katherine Michelmore, Stephanie Owen, and Shwetha Raghuraman. The power of certainty: Experimental evidence on the effective design of free tuition programs. No. w29864. National Bureau of Economic Research, 2022.

[17] Camacho, Adriana, Julián Messina, and Juan Uribe Barrera. "The expansion of higher education in Colombia: Bad students or bad programs?." Documento CEDE 2017-13 (2017).

[18] Carranza, Juan Esteban, and María Marta Ferreyra. "Increasing higher education access: Supply, sorting, and outcomes in Colombia." *Journal of Human Capital* 13.1 (2019): 95-136.

[19] Cattaneo, Matias D., Rocío Titiunik, and Gonzalo Vazquez-Bare. "Analysis of regression-discontinuity designs with multiple cutoffs or multiple scores." *The Stata Journal* 20.4 (2020): 866-891.

[20] Cattaneo, Matias D., and Rocio Titiunik. "Regression discontinuity designs." *Annual Review of Economics* 14 (2022): 821-851.

[21] Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. "A practical introduction to regression discontinuity designs: Extensions." arXiv preprint arXiv:2301.08958 (2023).

[22] Choi, Jin-young, and Myoung-jae Lee. "Complier and monotonicity for Fuzzy Multi-score Regression Discontinuity with partial effects." *Economics Letters* 228 (2023): 111169.

[23] Clark, Damon, and Paco Martorell. "The signaling value of a high school diploma." *Journal of Political Economy* 122.2 (2014): 282-318.

[24] Cortes, Patricia, Jessica Pan, Laura Pilossoph, Ernesto Reuben and Basit Zafar. "Gender Differences in Job Search and the Earnings Gap: Evidence from the Field and Lab" forthcoming, *Quarterly Journal of Economics*.

[25] Cohodes, Sarah R., and Joshua S. Goodman. "Merit aid, college quality, and college completion: Massachusetts' Adams scholarship as an in-kind subsidy." *American Economic Journal: Applied Economics* 6.4 (2014): 251-285.

[26] Dillon, Eleanor Wiske, and Jeffrey Andrew Smith. "Determinants of the match between student ability and college quality." *Journal of Labor Economics* 35.1 (2017): 45-66.

[27] Dillon, Eleanor Wiske, and Jeffrey Andrew Smith. "The consequences of academic match between students and colleges." *Journal of Human Resources* 55.3 (2020): 767-808.

[28] Dynarski, Susan, C. J. Libassi, Katherine Michelmore, and Stephanie Owen. "Closing the gap: The effect of reducing complexity and uncertainty in college pricing on the choices of low-income students." *American Economic Review* 111.6 (2021): 1721-1756.

[29] Eng, Amanda, and Jordan Matsudaira. "Pell grants and student success: Evidence from the universe of federal aid recipients." *Journal of Labor Economics* 39.S2 (2021): S413-S454.

[30] Ferreyra, María Marta, Ciro Avitabile, and Francisco Haimovich Paz. At a crossroads: higher education in Latin America and the Caribbean. World Bank Publications, 2017.

[31] Foote, Andrew, and Kevin M. Stange. Attrition from Administrative Data: Problems and Solutions with an Application to Postsecondary Education. No. w30232. National Bureau of Economic Research, 2022.

[32] Gill B., Lockwood J. R., Martorell F., Setodji C. M., Booker K. (2007). State and local implementation of the No Child Left Behind Act. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

[33] Goodman, Joshua, Michael Hurwitz, and Jonathan Smith. "Access to 4-year public colleges and degree completion." *Journal of Labor Economics* 35.3 (2017): 829-867.

[34] Hahn, Jinyong, Petra E. Todd, and Wilbert H. van der Klaauw. Evaluating the effect of an antidiscrimination law using a regression-discontinuity design. No. w7131. National Bureau of Economic Research, 1999.

[35] Heckman, James J., and Rodrigo Pinto. "Unordered monotonicity." *Econometrica* 86.1 (2018): 1-35.

[36] Hoxby, Caroline M., and Christopher Avery. "Low-income high-achieving students miss out on attending selective colleges." Brookings Papers on Economic Activity, Spring 10 (2013).

[37] Jacob, Brian A., and Lars Lefgren. "The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago." *Journal of Human Resources* 39.1 (2004): 50-79.

[38] Kozakowski, Whitney Catherine. Essays on higher education and inequality. Diss. Harvard University, 2020.

[39] Laajaj, Rachid, Andrés Moya, and Fabio Sánchez. "Equality of opportunity and human capital accumulation: Motivational effect of a nationwide scholarship in Colombia." *Journal of Development Economics* 154 (2022): 102754.

[40] Lee, David S., and David Card. "Regression discontinuity inference with specification error." *Journal of Econometrics* 142.2 (2008): 655-674.

[41] Lee, David S., and Thomas Lemieux. "Regression discontinuity designs in economics." *Journal of Economic Literature* 48.2 (2010): 281-355.

[42] Light, Audrey, and Wayne Strayer. "Determinants of college completion: School quality or student ability?." *Journal of Human Resources* (2000): 299-332.

[43] Londoño-Vélez, Juliana, Catherine Rodríguez, and Fabio Sánchez. "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 (2020): 193-227.

[44] Londoño-Vélez, Juliana, Catherine Rodríguez, and Fabio Sánchez, and Luis Esteban Alvarez. Financial Aid and Social Mobility: Evidence From Colombia's Ser Pilo Paga. Working paper (2023).

[45] Lovenheim, Michael F., and Jonathan Smith. Returns to different postsecondary investments: Institution type, academic programs, and credentials. No. w29933. National Bureau of Economic Research, 2022.

[46] MacLeod, W. Bentley, and Miguel Urquiola. "Reputation and school competition." *American Economic Review* 105.11 (2015): 3471-3488.

[47] MacLeod, W. Bentley, Evan Riehl, Juan E. Saavedra, and Miguel Urquiola. "The big sort: College reputation and labor market outcomes." *American Economic Journal: Applied Economics* 9.3 (2017): 223-261.

[48] Matsudaira, Jordan D. "Mandatory summer school and student achievement." *Journal of Econometrics* 142.2 (2008): 829-850.

[49] Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters. "The causal interpretation of two-stage least squares with multiple instrumental variables." *American Economic Review* 111.11 (2021): 3663-3698.

[50] Montalbán, José. "Countering moral hazard in higher education: The role of performance incentives in need-based grants." *Economic Journal* 133.649 (2023): 355-389.

[51] Moreira, Marcelo J. "A conditional likelihood ratio test for structural models." *Econometrica* 71.4 (2003): 1027-1048.

[52] Pallais, Amanda, and Sarah Turner. "Opportunities for low-income students at top colleges and universities: Policy initiatives and the distribution of students." *National Tax Journal* 59.2 (2006): 357-386.

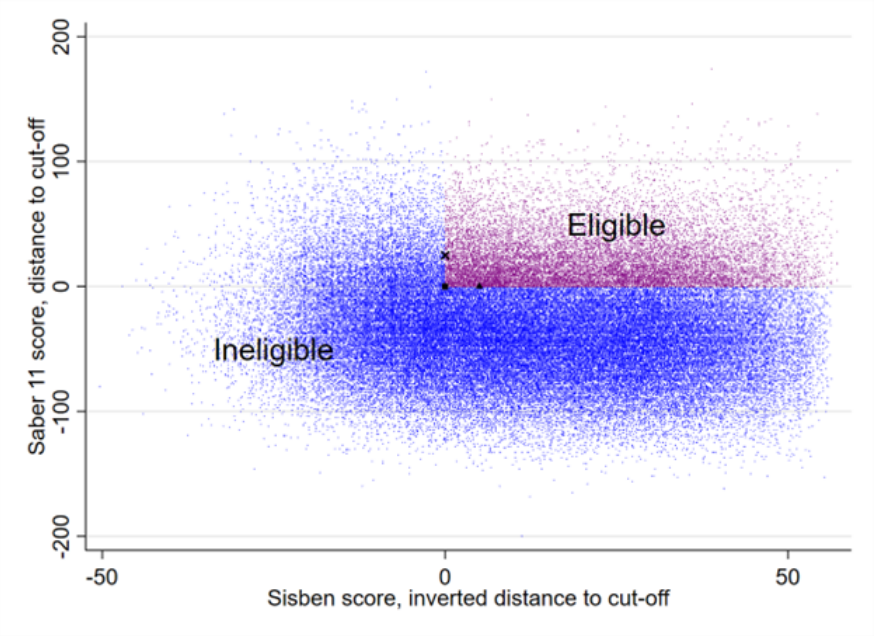
[53] Papay, John P., Richard J. Murnane, and John B. Willett. "The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts." *Educational Evaluation and Policy Analysis* 32.1 (2010): 5-23.

[54] Scott-Clayton, Judith. "On money and motivation: a quasi-experimental analysis of financial incentives for college achievement." *Journal of Human Resources* 46.3 (2011): 614-646.

[55] Stange, Kevin M. "An empirical investigation of the option value of college enrollment." *American Economic Journal: Applied Economics* 4.1 (2012): 49-84.

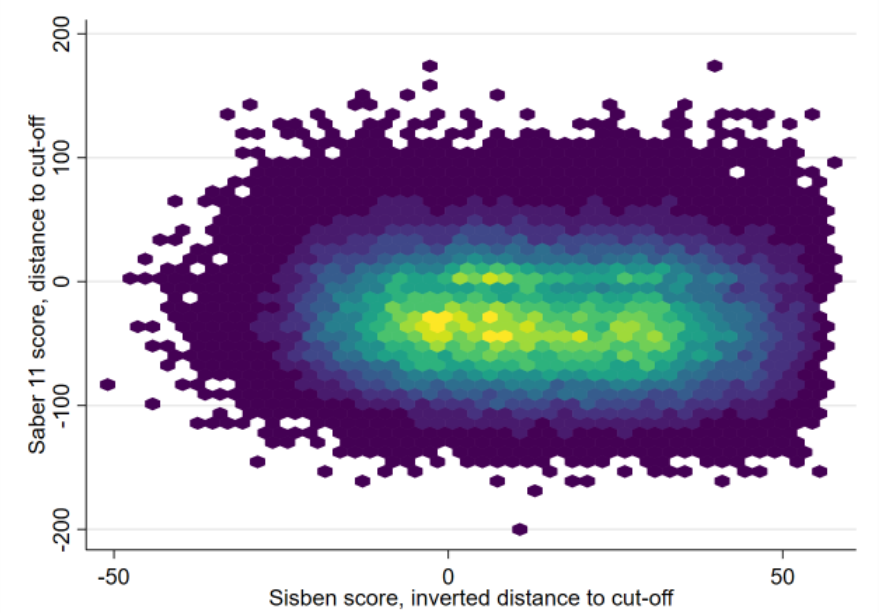
[56] Uquiola, Miguel. Markets, minds, and money: Why America leads the world in university research. Harvard University Press, 2020.

Figure 1: Scatterplot of Sisben and Saber 11 scores



Notes: Calculated by the authors. All scores have been normalized such that the cut-off is zero; the Sisben score has been inverted, so a score of 5 for example means the observation is 5 points below the qualification threshold. Dark dots indicate ineligible observations, while light colored dots indicate eligible observations for the SPP program. The square at (0,0) is the double-marginal point, the triangle at (0, 5) is the low wealth point, and the X at (25,0) is the high test performance point; see Section 4.1 for details.

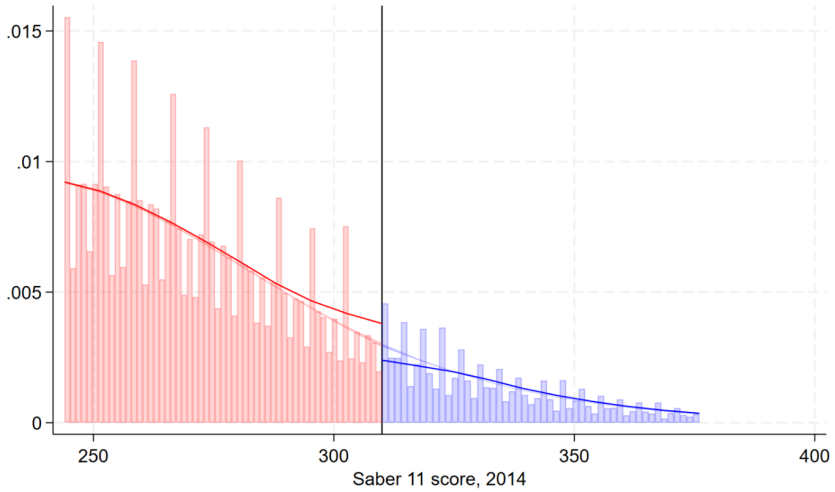
Figure 2: Hexagonal heat map of Sisben and Saber 11 scores



Notes: Calculated by the authors. All scores have been normalized such that the cut-off is zero; the Sisben score has been inverted, so a score of 5 for example means the observation is 5 points below the qualification threshold. The lightest hexes each represent approximately 0.335 percent of the sample, while each of the darkest hexes contain approximately 0.012 percent of the sample; see Section 4.1 for details.

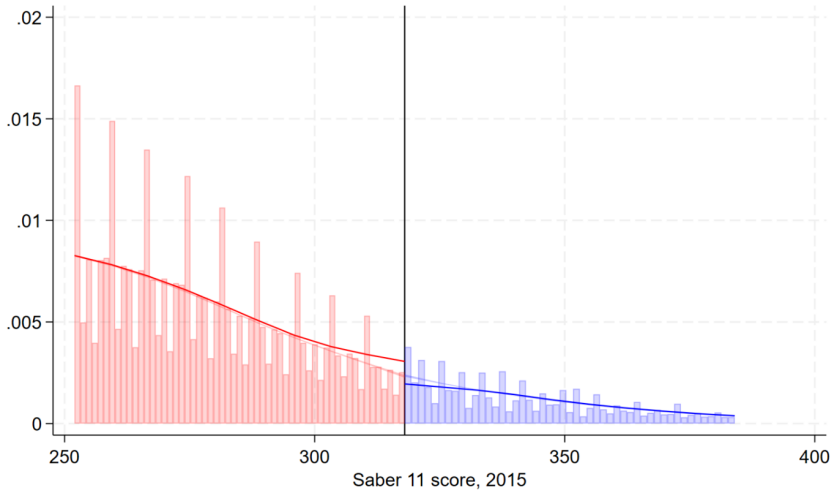
A. Appendix

Figure A.1: Density test, 2014 Saber 11 score



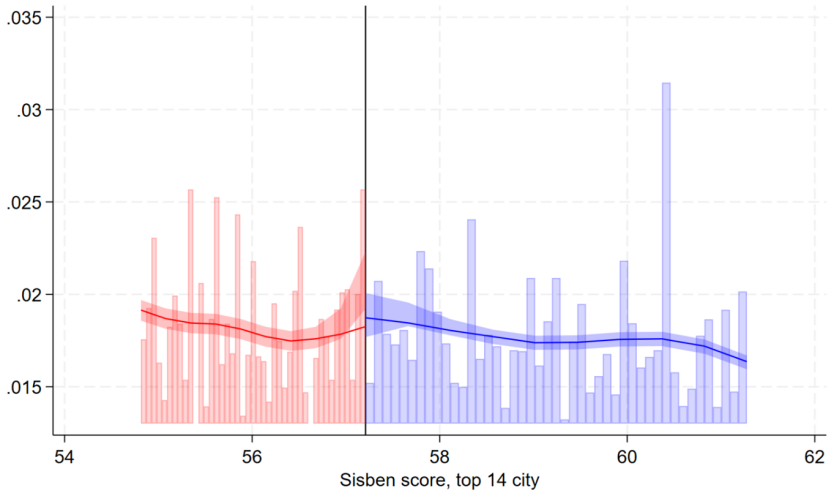
Notes: Calculated by the authors. Densities are calculated using a local linear specification and bias-corrected densities using a local quadratic specification. A triangular kernel is employed. Bias-corrected Simes p-values which correct for mass points are employed since the test score is a discrete running variable. We do not reject the null hypothesis of continuity at the cut-off (p-value = 0.918).

Figure A.2: Density test, 2015 Saber 11 score



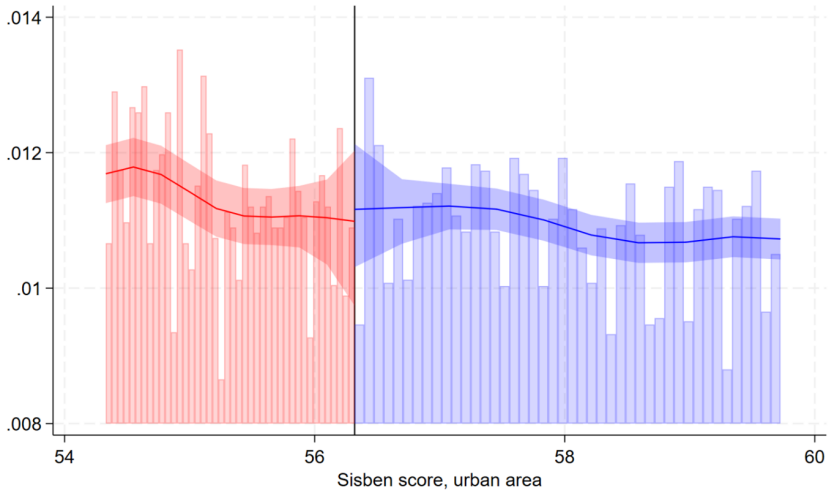
Notes: Calculated by the authors. Densities are calculated using a local linear specification, and bias-corrected densities using a local quadratic specification. A triangular kernel is employed. Bias-corrected Simes p-values which correct for mass points are employed since the test score is a discrete running variable. We do not reject the null hypothesis of continuity at the cut-off (p-value = 0.198).

Figure A.3: Density test, Sisben score, top 14 city



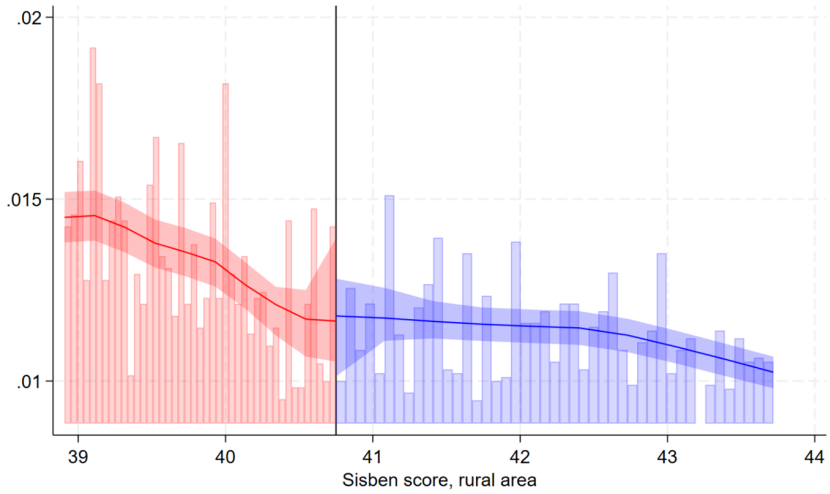
Notes: Calculated by the authors. Densities are calculated using a local linear specification, and bias-corrected densities using a local quadratic specification. A triangular kernel is employed. Bias-corrected Simes p-values which correct for mass points are employed since the Sisben score is a discrete running variable. We do not reject the null hypothesis of continuity at the cut-off (p-value = 0.198)

Figure A.4: Density test, Sisben score, other cities



Notes: Calculated by the authors. Densities are calculated using a local linear specification, and bias-corrected densities using a local quadratic specification. A triangular kernel is employed. Bias-corrected Simes p-values which correct for mass points are employed since the Sisben score is a discrete running variable. We do not reject the null hypothesis of continuity at the cut-off (p-value = 0.818).

Figure A.5: Density test, Sisben score, rural areas



Notes: Calculated by the authors. Densities are calculated using a local linear specification, and bias-corrected densities using a local quadratic specification. A triangular kernel is employed. Bias-corrected Simes p-values which correct for mass points are employed since the Sisben score is a discrete running variable. We do not reject the null hypothesis of continuity at the cut-off (p-value = 0.818).