Affirmative Action and Human Capital Accumulation: Evidence from Brazil

Marcos Hirai Catao

Professor Jason Baron, Faculty Advisor
Professor Duncan Thomas, Faculty Advisor

Honors Thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University.

Duke University
Durham, North Carolina
2024

Acknowledgements

In the course of crafting this thesis, my gratitude first and foremost extends to Professor Duncan Thomas and Professor Jason Baron. Their perceptive counsel not only steered this project to fruition but also significantly contributed to my development as a scholar during my undergraduate journey at Duke University. Their investment in mentoring me, both in this endeavor and in my overall growth as a researcher, has been invaluable.

Equally, my sincere appreciation goes to Professor Erica Field, Professor Matt Masten, and Professor Adam Rosen. Their guidance was instrumental in equipping me with the necessary skills to embark on this project.

I am thankful for the constructive feedback from Professor Michelle Connolly and my colleagues in the Honors Seminar. Furthermore, the support and input from the members of the Frankenberg-Thomas Lab, particularly the enlightening discussions with Peter Katz, Richard Lombardo, and Andres Santos, have significantly enhanced the quality of this project.

My profound gratitude is owed to Professor James Roberts and the Duke Economics Department for generously funding the data access crucial for my thesis. This support was pivotal in realizing the research.

A special acknowledgment is due to the staff at INEP, whose assistance with data-related queries was exceptionally helpful.

Lastly, but by no means least, my heartfelt thanks to my friends and family. Their unwavering support and understanding have been a source of constant strength and encouragement throughout my four years of study.

Abstract

In this study, I examine the effects of affirmative action (AA) policies on high school students' incentives to invest in human capital, focusing on the Brazilian Quotas Law (QL). This law mandates that federal higher education institutions reserve half of their seats for students from public high schools. Utilizing administrative data on schooling, college enrollment, and performance on standardized tests, I observe an increase in test scores among private high school students who attend public colleges. This increase corresponds with the reduction in available non-reserved seats. Conversely, no significant change is observed in the performance of public school students, despite a substantial increase in reserved seats, indicating a potential behavioral response. To estimate the effects of the policy, I analyze variations in policy exposure across regions and cohorts using difference-in-differences methods, which predominantly yield precisely estimated null results. Finally, I discuss potential reconciliations of these, proposing avenues for further research to explain the discrepancies.

JEL classification: I2, I23, I24

Keywords: Affirmative action; Higher Education; Human Capital

1 Introduction

Affirmative action (AA) is a major source of contention in the higher education policy debate around the world. These policies often aim to make college more accessible to historically disadvantaged groups, the target population. This is frequently achieved by explicitly reserving seats for the target population but can also be done less explicitly by prioritizing students of a certain disadvantaged group in the college admissions process. Higher education affirmative action policies have received a great deal of attention from behavioral and social scientists, including economists. Studies have looked extensively at the impacts of these policies on the choice of higher education programs, the quality of the student-college match and college graduation, among other outcomes. Recent literature has increasingly explored the association between affirmative action policies and pre-college human capital investment. The present study contributes to this growing area of research by investigating how large-scale, quota-based affirmative action policies affect the incentives for investing in human capital during high school. In particular, I analyze the effects of the Brazilian Quotas Law (QL) on the accumulation of human capital before college. I use administrative data on the universe of high school students in Brazil from 2007 to 2019, which allows me to observe cohorts that graduate high school before and after the policy. I identify treatment and control groups based on proximity to colleges affected by the QL and employ a difference-in-differences strategy to estimate causal effects of the policy.

In 2012, the Brazilian Government passed the Quotas Law (QL), which mandated that every federal university reserve 50 out of every 100 seats for students who attended public high schools, with 25 of these 50 seats further reserved for low-income students and 25 of the 50 seats reserved for under-represented minority (URM) students (Lei de Cotas, 2012). Brazil presents an interesting setting for the present study for three main reasons. First, university admissions decisions in Brazil are essentially based on a single test score, and the AA policy being studied established a clear fraction of quotas for the target population. Combined, these features of the setting make it easier for the student to assess the extent to which their chances of being admitted to each undergraduate program changed as a result of the policy.

Another reason for studying the Brazilian Quotas Law is that it is arguably much larger than the US affirmative action policies previously studied in the literature. Between 2012, the year the QL was implemented, and 2015, the first year when all reservations had to be implemented, the number of reserved seats for undergraduate students at federal colleges and universities each year increased approximately 80,000. For reference, in 2023, 70,753 new undergraduates enrolled in the University of California across all campuses and 51,913 new undergraduates enrolled in the University of Texas. Analytically, the scale of the policy presents both benefits and challenges. On the one hand, it marks a sharp change in the college admissions process in Brazil. On the other hand, the size of the program can make it difficult to characterize the heterogeneity of its impact. Despite the analytical challenges it poses, the QL's scale makes it more comparable to nationwide bans on AA than the more localized policies previously studied.

A third reason for choosing the Brazilian setting for the present study is data availability. Brazil possesses detailed administrative data that allows me to observe a wide range of educational outcomes. I use identified data from annual censuses of primary, secondary, and tertiary education institutions to observe students' educational trajectory, including which schools they attended, what year they started high school, what year they graduated (if they graduated at all), whether they enrolled in college, and, if they did, what program they enrolled in. Using identified data from the National High School Exam (ENEM), I also observe students' test-taking behavior and the test scores of those who choose to take the test. The sensitive nature of these data posed several challenges to the analysis. Obtaining access to the data required the submission of a project proposal to the Anísio Teixeira National Institute of Education Research (*Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*, also known as INEP). Once the project is approved, access to the data is only provided at one of INEP's data centers, including one in Brasilia. Despite the challenges of working with these data, they are essential for approach to answering the questions posed in the present study.

As previously mentioned, one notable feature of the Quotas Law is its scale: it was a countrywide policy that established uniform reservation rules across all federal institutions. However, this same feature complicates identification of causal effects of the policy as there is no obvious control group. To deal with this, I leverage the fact that the policy had heterogeneous effects across the country for two main reasons. First, most Brazilian universities do not provide housing to students and moving in order to attend college is costly. Second, several federal education institutions had already adopted affirmative action policies to some extent prior to the enactment of the Quotas Law. Consequently, it is plausible to assume that the Quotas Law's impact would be less pronounced in regions distant from federal colleges and universities or in areas where such institutions had pre-existing affirmative action policies. I combine this source of variation with comparisons across cohorts to estimate causal parameters, while also outlining the assumptions necessary for interpreting these parameters causally.

Using variations in exposure to the affirmative action policy and cohort comparisons within a difference-in-differences framework, I analyze the responses of high school students to the Quotas Law. In particular, I investigate the policy's effects on outcomes associated with human capital investment decisions, such as test scores, college enrollment, and timing of high school graduation. The methods employed in the present work are different from those used by notable empirical studies of affirmative action policies that estimate structural model of students' and colleges' preferences and simulate outcomes under counterfactual policies. Arcidiacono (2005) and Howell (2010) both use this approach and they both find that removing affirmative action would reduce the proportion of target students attending more selective schools but would have little impact on college attendance rates for target and non-target students.

Methodologically, this paper is similar to previous work that has used a difference-in-differences strategy to estimate the effect of affirmative action and similar changes to college admissions policies on college and labor market outcomes. Notable examples are Hinrichs (2012), Hinrichs (2014), Bleemer (2022), and Black et al (2023). The first three analyze the effects of affirmative action bans on the racial gap in college enrollment, graduation, and labor market earnings, while Black et al. analyze the effects of the Top Ten Percent policy in Texas on a similar range of outcomes. First, I investigate the response of students to the policy before college, an analysis not possible if the only outcomes used are college graduation and labor market earnings. Second, I use more granular variation in exposure to the policy in order to identify its effects. This allows me to separately identify causal effects for the target and non-target students while also restricting comparisons to be within-state.

The branch of the affirmative action literature that most closely relates to this project is that which analyzes the effects of Affirmative action on human capital investment decisions. A study conducted by Cotton et al. (2022) provides experimental evidence that an AA-like incentive structure can incentivize the target group to spend more time studying and perform better in exams. Studies conducted by Khanna (2020) and Akhtari et al (2020) both tackle questions similar to the ones I investigate. Khanna studies the effect of reservations on years of schooling in India using state-level variation as well as a regression discontinuity. The author finds that the policy reduced the gap in years of schooling between the target and non-target populations. Akhtari and coauthors focus on a Supreme Court decision that overturned a prior ban on race-based affirmative action in three states in the US. Using synthetic controls, the authors find that affirmative action led to an increase in the pre-college test scores of both minority and non-minority students, with a larger effect for minority students.

The present study complements the work of Khanna (2020) and Akhtari et al (2020) in important ways. First, the affirmative action policy in question is quota-based, which differentiates it from most recent affirmative action policies in the US. Second, the QL reserved university spots based on the type of school the student attended, not just race, ethnicity or caste, which makes the policy different from other AA policies studied in the literature. Finally, the source of identifying variation I employ allows me to estimate causal effects separately for the target and non-target population while restricting comparisons to be within-state. Previous studies have often used state-level variation to investigate the effects of affirmative action separately for the target and non-target population, and have only used more granular comparisons to estimate the effect of the policy on the gap between target and non-target students. Furthermore, the effects found using state-level variation in Akhtari et al (2020) are not found using the strategy I employ, which suggests that the results can vary depending on the institutional context and the source of variation used.

In the specific context of quota-based affirmative action policies in Brazil, economists have studied the effects on college graduation, earnings, and employment using data and policy changes from specific universities. A more recent set of studies investigates the effects of the affirmative action policy nationwide. Otero Barahona and Dobbin (2021) estimate a structural model of centralized college admissions system in Brazil and they find that the total loss in earnings by the non-target group is almost offset by the gains of the target population. Mello (2022) investigates the changes in the composition of the students who enroll in federal universities as a result of the QL. The author also shows that the QL induced a reduction in the average test score of incoming students, but that the reduction can be mostly explained by the change in composition. In a separate paper, Mello (2023) focuses on the choice of high school, showing that the QL induced students to switch from private to public schools at the start of high school. This study also provides suggestive evidence of a reduction in dropout rates and an increase in enrollment rates at federal colleges and universities.

While this paper differs substantially from the work of Otero Barahona and Dobbin (2021) both in terms of methodology and outcomes of interest, my contributions to the work of Mello (2022, 2023) are more nuanced. First, I refine the use of distance to affected colleges used in Mello (2023). While Mello (2023) measure exposure to the policy at the level of microregions (clusters of adjacent municipalities) with a sample of 50 microregions, I use variation in exposure to the policy at the school level for identification. This allows for within-state comparisons, reducing the scope for violations of the parallel trends assumption. Second, using this alternative source of variation, I analyze the effects on the educational outcomes of both public and private school students. Third, I provide evidence of the effects on students' test scores and their choices of college type and program. This complements the finding in Mello (2022) that test scores of the incoming students decreased as this negative effect may have been exacerbated or attenuated by changes in the behavior of students prior to attending college. Finally, I estimate the effects of the policy on overall college enrollment as well as on program choice for students who were in high school at the time the policy was announced. This complements the work of Mello (2023), which focuses on the effects on students who began high school after the policy change.

The rest of the paper proceeds as follows. In section 2, I provide more information on the Brazilian educational context and on the Quotas Law. In section 3, I describe the data that is used in the analysis. Section 4.1 provides details on how the outcomes are measured. Section 4.2 describes the relationship between distance to college and college enrollment in Brazil, which informs the identification strategy detailed in Section 4.3. Section 4.4 is devoted to estimation of the difference-in-differences model. Sections 5.1 and 5.2 respectively provide descriptive evidence on the trends in educational outcomes around the time of the policy and on the test scores of incoming college students. Model estimates are reported in section 5.3, and section 5.4 is devoted to discussion of the results. Section 6 concludes.

2 Setting

In Brazil, the majority of primary and secondary schools are public, while most post-secondary education institutions are private. Public universities can be managed by the Federal, State or Municipal government, with Federal institutions accounting for approximately 17% of the total undergraduate enrollment of approximately 6.5 million students in 2014 (Mello, 2022). That is, in 2014, approximately 1 million undergraduate students attended a federal university or technology institute. Federal universities usually focus on undergraduate education, and they can vary widely in size: some have an undergraduate population of less than 5,000 students while several others have more than 30,000 undergraduates. There is at least one of these universities in each state, and while many of them are located near the state's capital, several are located further away from urban areas. Federal and state universities are often regarded as high-quality, selective institutions. Admission into many state and federal universities is based solely on the score on the national examination known as ENEM. This exam is offered once per year and consists of four multiple choice sections with 45 items each and one essay. The exam grade can then be used by the students to apply to the universities through the centralized admissions system (SISU). Since private schools generally performed better on the national exam, for several years, most students at public universities had attended a private high school, even though private schools accounted for approximately 20 percent of high school students.

To make public universities more accessible to students from public schools and underrepresented minority students, the Brazilian government implemented an affirmative action policy in August 2012. The Law of Quotas reserved 50 out of every 100 slots at federal higher education institutions for students who attended public schools for all three years of high school. Additionally, out of the 50 reserved slots, 25 must be reserved for low-income students, and 25 for underrepresented minorities. The law stipulated that by 2015, approximately 142,000 seats at federal colleges and universities would be reserved for public high school students each year. It is key to note, however, that some of those seats had already been reserved as several institutions had already implemented quotas on their own prior to the QL. The total number of reserved seats in public colleges and universities increased by approximately 80,000 from the announcement of the policy in August 2012 to 2015. For a student to be considered for the reserved seats, said student must have attended a public school for all three years of high school. Therefore, students who were in a private high school at the time the policy was enacted would not be eligible for the quota if they switched to a public school. However, some private school students who were not yet in high school had large incentives to switch to public schools. Mello (2023) finds that in practice approximately 5% of private school students responded to the change in incentives and switched to public schools at the start of high school.

As previously mentioned, the Brazilian context is especially interesting for studying the effect of affirmative action on human capital prior to college. The policy environment differs from those of US affirmative action studies in key aspects. Admission into most Brazilian colleges is based solely on the national exam score, arguably making it easier for students to adjust their expectations in response to this policy. In addition, the QL is a quota-based nationwide affirmative action policy. This makes it different from most AA policies studied in the US, which are often more local in nature and which do not involve quotas — at least since quotas were deemed unconstitutional by the Supreme Court in 1978. These aspects, including the nationwide scope of the quota policy and reliance on standardized exam scores increase the difficulty of predicting the consequences of the QL just based on empirical evidence from the US. The availability of comprehensive administrative data on educational outcomes also permits a rigorous analysis of the effects of the policy.

3 Data

Investigating the relationship between affirmative action and human capital investment before college will require the combination of five administrative datasets. The first source of data is the yearly school census (CEB), which provides student-level information including age, gender, and school grade, as well as school-level information. The CEB is conducted every year and virtually all primary and secondary education institutions must provide data on all the students enrolled. The CEB allows one to track the educational history of the universe of students who are enrolled in a school in Brazil. With data from the CEBs from 2007 to 2019, it is possible to measure grade progression and retention for every student enrolled during that period.

In order to measure students' performance, I use microdata from the national high school examination (ENEM) for years 2009 through 2019. The exam is offered once per year and consists of one essay and four sections of 45 multiple choice questions: math, natural sciences, language, and humanities. Each test taker receives a section-specific numeric score based on which questions they got right. Test takers are identified by their CPF, which allows me to match students across the CEB and ENEM datasets.[1]

In addition to the CEB and the ENEM microdata, I will also use student-level data from the census of higher education (CES) for the period of 2009-2022. The census of higher education collects data like the ones collected by the CEB, but for all higher education institutions. The CES dataset also includes each student's CPF, which allows me to match students in the CES data to their respective records in the CEB and ENEM datasets.

Because I rely on a personal identifier to link students' records across the CEB, ENEM and CES datasets, the data I use in this project are not publicly available. Due to the sensitivity of the personal identifiers, access to these data is restricted, and the researcher must submit a project proposal to INEP, detailing the objective of the research, describing which data are being requested and explaining why using identified data is essential for achieving the objectives of the

[1] Not all students in the CEB dataset have a CPF. However, a CPF is required for taking ENEM, so it is possible to match the students in CEB who took the test. Measurement issues that arise due to the optional nature of the national exam are discussed further in section 4.1.

project. If the project is approved, the researcher may then access the requested data in one of the facilities approved by INEP, one of which is in INEP's headquarters in Brasilia.

Data on the adoption of quotas before and after the QL by each public college and university is obtained from the dataset compiled by Mello (2022). This dataset contains the number of seats designated for each affirmative action category at every public higher education institution from 2010 to 2015.

Using the school-level (including zip code and address) and college-level information (including college campus name and municipality name) from the CEB and CES, respectively, I geocoded each high school and each college in the sample. Because the geocoding was considerably less precise for schools located in the Amazon, high schools located in the Amazon were removed from the sample. This is likely not to pose a problem in terms of statistical power as the Amazon is sparsely populated and the remainder of the country corresponds to 87% of the population. Additionally, the large distances between schools and colleges in the Amazon may pose a problem for the identification strategy described in section 4.3.

In order to conduct the analysis, I first restrict the sample to students in the CEB that attended their first year of high school at any time between 2008 and 2012. To avoid endogeneity issues that may arise if students transfer across schools during high school, I assign students to the school where they first enrolled in the year when they started high school.[2] The students are then assigned to a cohort based on the year they first enrolled in high school. A student who enrolled in high school for the first time in year t and progresses smoothly through high school

[2] Fewer than 0.8 percent of students are simultaneously enrolled in more than one school in their first year of high school. For these students, I randomly assign them to one of the schools.

(that is, the student is not held back and does not leave school for an extended period of time) would graduate high school at the end of year $t + 2$. Thus, for the remainder of the paper, I will refer to the cohort of students that first enrolled in high school in year t as the graduating cohort of year $t + 2$, even though cohort assignment is based solely on the year of enrollment. Finally, I restrict the sample to the graduating cohorts of 2010 through 2015, as that permits the measurement of the outcomes of interest.

Summary statistics for the 2010 graduating cohort in the analysis sample are reported in Table 1. Columns 1 and 2 report the means of the variables separately for public and private schools, respectively, and the overall means are reported in column 3. Table 1 demonstrates the differences between private and public schools. First, the majority of students in the 2010 cohort attend a public school—public schools account for 87 percent of total high school enrollment. Panel A in Table 1 shows that the composition of the cohort differs across public and private schools. Students who attend public high schools are more likely to be racial minorities and are older on average—at the time they first enroll in high school, public school students are on average 1.9 years older than their private school counterparts. Differences in the age of students who are in the same cohort likely occur due to a combination of two factors: starting school late and being held back. The latter is a common occurrence—according to the 2018 PISA report (OECD, 2020), 34% of 15-year-olds in Brazil had been held back at least once. Panel A also shows that relative to students who go to a public school, private school students are approximately 50 percent more likely to finish high school on time (in 3 years),[3] and that the difference is similar when we consider graduation within 4 or 6 years. Panels B and C report

[3] In Brazil, there are three grades in high school, but because the school year starts in January, someone who starts high school in, say, 2005, will graduate in the end of 2007.

college enrollment outcomes and standardized test scores measured within 3 and 6 years of high school enrollment, respectively.[4] Relative to public schools, private schools show higher enrollment rates in both private and public colleges within 3 and 6 years of starting high school. Private school students also take ENEM at almost double the rate of public school students, and the average score of the top performing students in private schools is approximately one standard deviation higher than that of their counterparts in public schools for each section of the exam.

4 Methodology

Because the Quotas Law was a national policy, with all units being treated at once, there is no clear control group that could be used to remove time trends. For this reason, I leverage heterogeneity in exposure to the policy along two dimensions: distance to the nearest federal higher education institution, and fraction of seats at the nearest federal college or university. In this section, I explain in more detail how these sources of heterogeneity will be used to recover parameters that can be interpreted as causal. I first describe the measurement of the outcomes of interest. Then, I provide descriptive evidence on the relationship between enrollment in higher education and distance to the nearest college or university. Finally, I explain how that relationship will be used for identification, and describe the procedure for estimating the difference-in-differences model.

4.1 Measurement

The main goal of the present study is to understand the effect of affirmative action on human capital accumulation of high school students. One set of outcomes that is of interest

[4] To account for variation in the mean and dispersion of test score distributions, I use test-year-specific z-scores for each subject. For students who take the exam more than once within 3 or 6 years of high school enrollment, I take the highest z-score.

pertains to educational attainment. If the policy affected human capital accumulation, then the effect may be captured by the likelihood that students complete high school. When measuring high school completion in Brazil, it is key to consider the timing of high school completion since a substantial fraction of high school graduates take longer than the standard 3 years to complete high school. Table 2 reports the distribution of the year when students in the 2010 graduating cohort complete high school, enroll in college, and enroll in a public college. This table demonstrates that approximately 16 percent of the students who graduated high school did not do so on time. Fewer than 2 percent of the students who graduate high school do so in more than 6 years. For that reason, high school completion is measured within 3 and 6 years of enrollment. The first is meant to capture timely high school completion while the latter measures whether the student graduates from high school at all.

College enrollment and program choice may also capture changes in human capital accumulation. Interpreting the effects of the QL on college-enrollment-related outcomes can be difficult because it can be difficult to untangle changes that are mechanical from ones that are behavioral. Nonetheless, college and program choice are crucial aspects of the response to the policy and explaining changes in the human capital accumulation will likely require an investigation of the effects of the QL on these outcomes. Thus, I measure enrollment in college, enrollment in a public college, enrollment in a STEM program, and enrollment in a STEM program in a public college. These variables are also measured twice, once within 3 years of high school enrollment, and once within 6 years. The timing of enrollment in college may pose a measurement problem since almost 20 percent of the students who enrolled in college or in a public college did so more than 6 years after they enrolled in high school. The choice of limiting the measurement to 6 years was made mainly due to data limitations. The CES dataset only included years 2009 through 2019, so in order to measure college enrollment for the cohort that enrolled in 2013, the measurement period had to be limited to 6 years from the start of high school.[5]

Another way to measure human capital accumulation is to use test scores, and in the present project, ENEM is the main source of test score outcomes. For the purposes of the present project, an ideal test would be administered to all students, or at least to a representative sample of each cohort. However, ENEM is an optional test, and students can use their score on the test to apply to college. Consequently, the decision to take the exam is likely to be affected by the policy. Thus, an analysis of the effect of the policy on the scores of the students who decide to take the test may be contaminated by selection bias. While I show in section 5.3 that the policy did not affect the probability of taking the exam, I propose an assumption that allows me to obtain a measure of test scores that is less likely to be affected by selection. In particular, I assume that the students who do not take ENEM would have been at the bottom of the test-score distribution within their school and cohort. Under this assumption, we can observe the top quantiles of the distribution of potential test scores. That is, if we think of test scores as a potential outcome that is only observed for the students who choose to take the test, then this assumption implies that the test scores that we do observe correspond to the top quantiles of the within-school-and-cohort potential test score distribution. More precisely, under this assumption, if in a school s and cohort t a fraction α of students took the test, then the top α percentiles of the distribution of potential test scores in school s and cohort t are identified. The assumption may be justified by the fact that there is a cost to taking the exam, not only because there is a fee,

[5] I have since then received access to CES data for the 2020-2022 period, so the interval of measurement can be expanded in a future iteration.

but also because taking the exam requires up to 10 hours over two weekends. Thus, students who do not think they can perform as well as their peers and, as a result, use the score to get into college, have fewer incentives to take the exam. For the school-cohort cells where at least 10 percent of students took the exam, we compute the average score for the top 10 percent performing students in each section of the exam. This allows me to maintain approximately 92 percent of the original school-cohort cells for each section of the exam. I also compute the analogous within-school-and-cohort average for the top 20 and 30 percent of students, but these restrict the sample to 80 and 60 percent of the original school-cohort cells, respectively.

4.2 Distance to College and College Attendance

Two features of the setting are crucial for identification of causal parameters in this study. First, college attendance in Brazil is correlated with distance to the nearest college. Second, some federal colleges and universities had already reserved seats for public school students prior to the implementation of the policy.

Distance to college can pose a major challenge to college attendance in Brazil as moving away from home to attend college may substantially increase the cost of attendance. According to the CES, approximately 50 percent of the students who enrolled in college in 2010 did so in the same municipality where they were born, and that approximately 75 percent attended college within 50km from the municipality where they were born. Considering that students may have moved since they were born, it is possible that an even higher fraction of students attended a college close to the high school they attended.

To further investigate the relationship between distance to a college and the probability of going to college, I restrict the sample to students from the 2010 graduating cohort and estimate the following model

$$y_i = f(z_i) + g(w_i) + \epsilon_i,$$

where y_i is an indicator variable that takes value 1 if student i enrolled in college within 6 years of enrolling in high school, z_i is the sixth root of the distance between student i's high school and the nearest college that offers at least 200 seats[6], and w_i is a set of control variables, which includes state fixed effects, the average income and population density of the school's census tract, and the number of high school students enrolled within a 30km radius around the school. In this model, I assume that the function $g(\cdot)$ is additively separable in each of the components of w_i, including the fixed effects. However, the shape of each of the continuous components is allowed to be flexible, with each component being estimated using a nonparametric series estimator. The function $f(\cdot)$ is also estimated nonparametric using a series estimator. Note that since state fixed effects are included in the regression, the intercept of $f(\cdot)$ is not identified. This does not pose a problem since only the gradient of $f(\cdot)$ is of interest.

The negatively sloped relationship between y_i and z_i is clear in figure 1, which plots $f(\cdot)$ on the right and the density of z_i on the left. Interpreting the shape of $f(z_i)$ for $z_i \leq 1$ is difficult since these schools are within less than one kilometer away from a public university and are

[6] The choice of only considering colleges that are at least of a certain size was made because of higher education institutions where few students enroll each year, as these are likely to not offer many programs and consequently probably do not draw interest from many high school students. Analogous variables that only consider colleges with at least 150, 300 and 400 seats were computed and the correlation between each of these variables and d_i ranged between 85 and 92, which suggests that the results would have been similar if any of these variables was used insead of d_i.

likely differ substantially from the other schools. However, consider z_i in the range [1,1.8], which corresponds to a distance to a distance to the nearest college ranging from 1 to approximately 35 kilometers. In that range the expected probability that the student attend college within 6 years of starting high school goes down by approximately 7 percentage points, which corresponds to approximately 30 percent of the sample mean.

4.3 Identification

Identification of causal effects of the Quotas Law relies on two features of the setting. First, as established in section 4.2, college attendance is strongly correlated with distance to the nearest college, and a large fraction of college students do not move far to attend college. Second, several federal colleges and universities already had some fraction of seats reserved for public high school students when this policy was implemented. Thus, exposure to the policy varied widely across the country. Intuitively, if a large number of seats was reserved in a given college because of the policy, then the high school students in a nearby area are more likely to respond. On the other hand, students who live far enough away from the affected college are considerably less likely to change their behavior because of the barrier imposed by the distance to that college. The idea behind the identification strategy is to use the students who live far from an affected college as the control group, so that trends in their outcomes reflect changes that the students who live close to affected colleges would have experienced in the absence of the policy. This is similar to a shift-share instrumental variable strategy, where exposure to a plausibly exogenous shock is mediated by a predetermined characteristic.

I first construct a measure of exposure. For a student i and a college campus c, let d_{ic} be the distance between the student's high school and the college campus and let ql_c be the number of seats reserved by the QL in c. First, I compute the smallest distance r_i such that 200 seats are reserved by the policy within a radius r_i of student i's high school.[7] To account for the fact that the number of seats reserved in colleges and universities within the radius r_i can vary, I then compute the exposure variable s_i which is the average of d_{ic} for colleges within the radius r_i, weighted by the number of reserved seats at the given college. Then, since the relationship between distance to college and attendance becomes relatively flat after $30km$, I define a binary variable D_i which takes value 1 if $s_i \leq 30km$, so that $D_i = 1$ denotes the exposed group.

Some additional notation is required to define the target parameters, formalize the identification strategy and explain the estimation procedure. First, let X_i denote a vector of covariates which includes state indicator variables and continuous covariates that measure the demographic characteristics of the census tract where student i's high school is located. Let K_i denote the graduating cohort of student i, relative to the year of 2011, so that $K_i = 0$ indicates that student i belongs to the graduating cohort of 2011, $K_i = 1$ for students in the graduating cohort of 2012, and so on. Let $Post_i$ be a binary random variable such that $Post_i = 1$ if $K_i > 0$ and $Post_i = 0$ otherwise. For a given outcome of interest Y_i, define the potential outcomes $Y_i(0), Y_i(1)$, where $Y_i(0)$ is the outcomes of individual i in the absence of the policy and $Y_i(1)$ is the outcome in the presence of the policy. Then, the observed outcome is written as $Y_i = Post_i \times D_i \times Y_i(1) + (1 - Post_i \times D_i) \times Y_i(0)$. Finally, I define the target parameter to be an average treatment effect on the treated (ATT); formally, this class of parameters is defined by

$$ATT_t = E(Y_i(1) - Y_i(0) | K_i = t, D_i = 1). \tag{4.1}$$

[7] The choice of 200 seats may seem arbitrary, but I compute the same exposure variable using 100, 300 and 350 seats instead and the correlation between all of them is above 0.90.

One of the key assumptions required for identifying the parameter defined in 4.1 is embedded in the definition of the observed outcome. In particular, the definition of the relationship between the observed and potential outcomes implies that $Y_i(0)$ is observed for students in graduating cohorts with $K_i \leq 0$. This assumption implies that there is no anticipation. It would be violated, for instance, if students who were in the 2011 graduating cohort changed their behavior while still in high school in anticipation of the policy change. The QL was approved in the second semester of 2012, approximately three months after the Brazilian Supreme Court ruled that racial quotas were legal. Thus, most of the 2011 graduating cohort had graduated high school prior to the approval of the policy, so even if they anticipated the policy by over a year, they would have had little time to adjust to it while still in high school. While a non-trivial fraction of the 2011 graduating cohort did not graduate on time, students from that cohort who were still in high school at the time the law was passed are likely not to be at the margin of going or not going to college. Therefore, these students are less likely to be the ones that would be affected by the policy.

Another key assumption is that conditional on the covariates X_i and in the absence of the treatment, the difference between the average outcomes of post-policy and pre-policy cohorts would have been the same across exposed and non-exposed groups. This conditional parallel-trends assumption can be stated formally as

$$\begin{aligned} &E(Y_i(0)|X_i, K_i = t, D_i = 1) - E(Y_i(0)|X_i, K_i = 0, D_i = 1) \\ &= E(Y_i(0)|X_i, K_i = t, D_i = 0) - E(Y_i(0)|X_i, K_i = 0, D_i = 0). \end{aligned} \tag{4.2}$$

This assumption seems reasonable in the present context because of exposure to the policy is a relatively complex function of the distribution of federal universities across the country as well as of pre-existing reservations. However, the assumption would be violated if the exposed regions were also areas where the supply of public college seats was growing at a faster rate. As discussed further in appendix A, the rate of growth of the supply of public college seats seems to be heterogeneous across the country, but most of the difference seems to be accounted for by the size of the market for college seats, which I control for in X_i.

4.4 Estimation

Under the identifying assumptions laid out in the previous section, it can be shown that the ATT_t parameters take the following form

$$ATT_t = E\left[m_{1t}(X_i) - m_{10}(X_i) - \left(m_{0t}(X_i) - m_{00}(X_i)\right)|D_i = 1, K_i = t\right], \tag{4.3}$$

where $m_{gt}(x) = E(Y_i|X_i = x, K_i = t, D_i = g)$ is the conditional mean function. This identification result leads naturally to the two-step outcome-regression estimator for the parameters of interest (see Callaway and Sant'anna (2021) or Sant'anna and Zhao (2020) for more details on estimation and inference). In the first step, I assume a parametric functional form for each $m_{gt}(x)$. In the main specification, each $m_{gt}(x)$ is assumed to be linear in the parameters, and I allow for state-specific intercepts as well as state-specific coefficients for the continuous variables. The parameters are then estimated using least-squares to obtain estimates $\widehat{m}_{gt}(x)$ of the functions $m_{gt}(x)$. In the second step, I obtain the fitted values $\widehat{m}_i^{gt} = \widehat{m}_{gt}(X_i)$. I then replace the $m_{gt}(X_i)$'s in 4.3 with the analogous fitted value and compute the sample analog of 4.3 to obtain an estimator $\widehat{ATT}_t$ for ATT_t.

It is important to note that since all the covariates X_i are constant within school, we can equivalently estimate the model by taking the school-by-cohort average of the desired outcome, and then weighting each school-by-cohort cell by the number of students. Given the large number of observations in the data, this fact becomes crucial for estimation as it substantially reduces the computational resources necessary to estimate the model and perform inference.

5 Results

I begin this section by reporting the general trends in outcomes of public and private school students around the time when the policy was implemented. I then provide evidence on the distribution of ENEM scores of students who enrolled in public colleges for each year and separately by the type of high school attended. The distributions suggest that public high students' test scores improved in response to the policy. On the other hand, difference-in-differences mostly yield precisely determined null effects. I propose two possible explanations for these results. One pertains to the use of distance as a source of differential exposure to the policy while the other hinges on heterogeneous treatment effects and the possibility that only a small fraction of students respond.

5.1 Trends in High School Graduation and Test Scores

Figure 2 displays the rate of on-time high school graduation separately for public and private high schools over the period of 2010 to 2015. To facilitate visualization, the values plotted are relative to the respective 2010 means, which are also reported in the figure. The figure shows an increase in the probability that students graduate on time for both public and private high schools. While the percentage point increase is similar across public and private schools, it must be noted that because of the difference in the original level, the increase over this period was 13% for public schools and 5% for private schools.

To contextualize the difference-in-differences estimates, it is also helpful to document the trends in the average test scores of the top students within each school evolved over the period of the analysis. To construct figure 3, I first compute, for each school-cohort combination, the average test score of the top 10% of students. Under the assumption that the students who do not take the test would have been at the bottom of the distribution of their respective schools, this average is identified. Then, I take a weighted average of these quantities by cohort and type of school, where the weights are given by the number of students in each school-cohort cell. In general, these figures show a large increase in the test scores of both public and private school students, with larger increases in the test scores of public-school students if we adjust for their respective 2010 standard deviations. The one exception to this is the Portuguese scores of private-school students, which did not change much over this period.

5.2 Test Scores of incoming college students

One way of investigating the effects of the policy is to look at the distribution of test scores for students who enroll in a public college in each year over from 2010 to 2015. The Quotas Law reserved over 100,000 college seats for public high school students, while reducing the number of seats available for private high school students in public colleges and universities. To investigate how this may have changed the profile of admitted students, I report the test scores of the who enrolled in a public college in each year from 2012 to 2015.

Public colleges and universities became more selective for students who attended a private high school, so in the absence of a behavioral response by these students, we would expect the test scores of those who enroll to increase. Figure 4 documents precisely this pattern. Each panel in figure 4 displays the empirical CDF of ENEM test scores for private high school students that were admitted into a public college or university for each year from 2012 to 2015. I exclude 2010 and 2011 because colleges only had to introduce quotas starting in 2013, so 2012 provides a pre-policy period. The figure shows that in general the 2012 test-score distribution in each subject is almost dominated by the distribution of scores in the following years. The pattern is especially notable in panel A, which plots the distribution of math test scores. Panel A documents a large rightwards shift in the distribution of test scores, especially at the top of the distribution, with the 80^{th} percentile increasing by approximately half a standard deviation from 2012 to 2015.

In contrast, the increase in the number of reserved seats makes public colleges and universities more accessible to public school students. Thus, in the absence of a behavioral response, one would expect the test scores of public high school students who enroll in a public college to decrease over from 2012 to 2015. However, this is not what we observe. Figure 5 plots the distribution of scores in each subsection of ENEM for students who attended public schools and enrolled in public colleges. One possible explanation is that public high school students who were enrolling in public colleges and universities after the policy perform just as well as the ones that did so prior to the policy. An alternative explanation is that students changed their behavior after the policy was implemented. In particular, it is possible that public high school students' scores improved enough to offset the expected left-ward shift that is expected if colleges become less selective for students who went to public high schools. In order to try to distinguish between these explanations, I report the test scores of public school students who enrolled in federal colleges and universities that were not affected by the policy because they had already reserved 50% of their seats for public high school students prior to the policy. In the absence of a behavioral response, the distribution of test scores of the public school students who enroll in these non-affected institutions should remain stable from 2012 to 2015. However, as figure 6 shows, the scores in each of the sections of ENEM increases for enrolled public high school students in these universities. This further suggests that public high school students' test scores increase after the policy.

5.3 Difference-in-Differences Estimates

While the evidence presented in section 5.2 suggests that the policy had a positive impact on the test scores of public high school students, there still are some alternative explanations for the observed patterns. For instance, changes in the choice of college and the composition of applicants may explain the increase in test-scores reported in figure 6. To further investigate the effects of the policy, I report estimates of the ATT_t parameters defined in 4.1. I estimate these parameters separately for private and public schools and for each graduating cohort from 2011 to 2015, inclusive. The estimates for the 2011 are placebo test of pre-trends and use the 2010 cohort as the comparison, while all other estimates use 2011 as the pre-treatment cohort. It is key to note that the students in the 2015 graduating cohort entered high school in the beginning of 2013, a few months after the Quotas Law was passed. Mello (2023) provides evidence that students changed their choice of high school in response to the policy, which would bias the estimates for the 2015 cohort. I still report the estimates for the 2015 cohort because Mello (2023) estimates the effect of the policy on the probability of switching from private to public schools for the 2015 to be approximately 3 percentage points, so sorting is not as problematic in 2015 as it is in the following years. Additionally, Mello (2023) finds evidence that the students who switched were positively selected relative to the peers in the destination school, so the estimates for the 2015 cohort may serve as an upper bound on the true estimate.

The estimated effects for private school students are generally not distinguishable from zero, although several of the estimates are noisy. Table 3 reports the effects on the probability of graduating high school on time and graduating high school in in 6 years. While none of the estimates are statistically significant, they are also generally noisy. For instance, for each of the estimates, we would fail to reject a positive or negative effect of 8 percentage points, which would correspond to more than 10 percent of the pre-policy mean. Table 4 reports the effects on enrollment in any college, in a private college, in a public college and in a STEM program. Again, I find no evidence that the program affected these outcomes for private-school students, but the standard errors are large, so I also cannot reject a large positive or negative effect. Finally, table 5 shows the estimated effect on the probability of taking the national exam and on the scores of the top performers. I find no evidence that the policy impacted the test-taking behavior of private-high school students or the test scores of the top-performing students. However, the confidence intervals are large and large positive and negative effects cannot be rejected.

Relative to the estimates for private high schools, the ones for public high schools are more precise. Table 6 reports the estimated effects on the graduation rates of public high school students. Columns 1 and 2 report the effects on the probability of graduating high school on time and in 6 years, respectively. Across all three columns, the estimated effects on high school graduation are essentially null. While the estimated effect on graduating on time is statistically significant for the 2013 and 2015 cohorts, these effects would likely become insignificant after adjusting for multiple testing. Moreover, the estimates for the 2015 cohort may suffer from selection bias, as previously noted, and the estimated effect of 3.2 percentage points for the 2013 cohort is relatively small, corresponding to approximately 8% of pre-policy mean of the exposed group.

Given the null effects on high school graduation, one would expect the estimated impact on other educational outcomes to be close to zero as well. Table 7 reports the estimated ATT_t's for college/university enrollment outcomes. Columns 1, 2 and 3 show precisely estimated null effects on enrollment in any college, in a private college and in a public college, respectively. Column 4 shows that the effect on enrolling in a STEM program is also a precisely estimated zero, suggesting that there was little effect on the choice of college programs. Table 8 reports the effects on test-taking behavior and test scores. The effect on the probability of taking the national exam is reported in column 1, and they show precisely estimated null effects. Columns 2, 3, 4 and 5 report the effects on the average test score of the top students within each school. The estimates are almost all statistically insignificant and small, with most of them contained in the 0 to 0.04 standard-deviation range. While the effect on math scores for the 2015 cohort is significant at the 5% level, adjustments for multiple testing would likely render this estimate insignificant. In addition, this estimate may be biased due to endogenous selection. Overall, these estimates suggest that the policy had little to no effect on the average outcomes of the treated group.

5.4 Discussion

Sections 5.2 and 5.3 present seemingly contradictory results on the effects of the QL on the outcomes of public high school students. The distribution of test scores of public high school students who enrolled in public colleges suggest that the policy increased the test scores of high school students. However, difference-in-differences estimates yield precisely estimated null effects, which are also different from previous in the existing literature that find positive effects on student outcomes. In this section, I propose two potential explanations for these results and suggest ways of testing these hypotheses as possible directions of future research.

One possible explanation for these results is that using distance to affected colleges and universities as a source of identifying variation leads to the estimation of parameters that are different from the true ATT_t. To see how that can happen, recall that the assumed relationship between the potential outcomes and the observed outcomes was defined to be $Y_i = Post_i \times D_i \times Y_i(1) + (1 - Post_i \times D_i) \times Y_i(0)$. As discussed before, this definition introduces a no-anticipation assumption. However, it also introduces a more subtle assumption. It implies that the observed outcome for the non-exposed group corresponds to the outcome under no treatment, so the distribution of outcomes observed in non-exposed areas is the same that would have been observed absent the policy. This essentially means that the policy had no effect on the non-exposed areas, which may not be the case. To check how the estimand can be interpreted if this assumption is relaxed, I instead assume that the observed outcome is $\tilde{Y}_i = Post_i \times Y_i(1) + (1 - Post_i) \times Y_i(0)$. Then, I define the conditional mean function $\widetilde{m}_{gt}(x) = E(Y_i|X_i = x, K_i = t, D_i = g)$ as well as the conditional average treatment effect on the treated,

$$ATT_t(X_i) = E(Y_i(1) - Y_i(0)|K_i = t, D_i = 1, X_i),$$

and the conditional average treatment effect on the untreated

$$ATU_t(X_i) = E(Y_i(1) - Y_i(0)|K_i = t, D_i = 0, X_i).$$

Under the original identifying assumptions, but with the new definition of the observed treatment, we obtain

$$\begin{aligned} &E\left[\widetilde{m}_{1t}(X_i) - \widetilde{m}_{10}(X_i) - \left(\widetilde{m}_{0t}(X_i) - \widetilde{m}_{00}(X_i)\right)|D_i = 1, K_i = t\right] \\ &= E[ATT_t(X_i) - ATU_t(X_i)|D_i = 1, K_i = t], \end{aligned} \tag{5.1}$$

where $E\left[\widetilde{m}_{1t}(X_i) - \widetilde{m}_{10}(X_i) - \left(\widetilde{m}_{0t}(X_i) - \widetilde{m}_{00}(X_i)\right)|D_i = 1, K_i = t\right]$ is the object being estimated if the observed outcome is $\tilde{Y}_i$ instead of Y_i. The result 5.1 implies that the estimand can be interpreted as a weighted average of the contrast between the conditional treatment effect on the treated and the conditional treatment effect on the untreated. Under the assumption that the distribution of outcomes in the non-exposed regions would have been the same in the presence and in the absence of the policy, we would have $ATU_t(X_i) = 0$, and the estimand would again coincide with ATT_t. Thus, if the model is misspecified, then my estimates will generally be inconsistent for ATT_t. In particular, if we believe that the effect of the policy would have the same direction but would be smaller in magnitude in the non-exposed regions, the result above shows that the estimates can be biased towards zero. To test this, one can try estimate the treatment effects separately for the nontreated areas, which can be done by comparing the outcomes of students of different races before and after the policy. Among the seats that were reserved for students from public schools, some were further reserved for racial minority students. Thus, if the policy affected students from public schools in non-exposed areas, then we may expect the effect to be larger for minority students.

Another possible explanation is that there is substantial treatment effect heterogeneity that is not captured by the research design employed. It is possible that the policy had a large effect on the outcomes of students who were on the margin between going to college and not going to college (marginal students), but that there was no effect on the outcomes of other students (non-marginal students). Then, the estimated ATT_t's would be a weighted average of the effects for marginal and non-marginal students. Since the policy reserved approximately 150,000 seats per year and there are over 2 million public school students in each cohort, marginal students would represent a small fraction of all students, so the average effect for marginal students would receive a small weight in the ATT_t's. Nonetheless, the scores of incoming students would have increased because the students at the margin respond positively to the policy. Testing this hypothesis may require identifying students who are likely to be at the margin. One possible way of doing this would be to use the demographic characteristics of students to predict what their exam scores would have been in the absence of the policy, and then compare those scores to the distribution of scores of students who enrolled in nearby affected colleges prior to the policy. Once students who are likely to be on the margin are identified, conditional average treatment-on-the-treated parameters can be estimated separately for marginal and non-marginal students to test the hypothesis.

6 Conclusion

The Brazilian Quotas Law was a large quota-based affirmative action policy. It reserved thousands of seats for public high school students in some of the most highly regarded colleges and universities in Brazil. In the present work, I assess the effects of this policy on human capital accumulation of public and private high school students by examining the distribution of test scores of students who enrolled in college around the time the policy was implemented, and by using heterogeneity in the exposure to the policy in a difference-in-differences framework. The estimated effects for private school students are imprecise, which makes it difficult to draw conclusions from them. Evidence from test-score distributions suggest that the policy had a positive impact on the exam scores of public high school students. However, difference-in-difference estimates yield precisely determined null effects. While I propose potential explanations for these seemingly incompatible pieces of evidence, more research is required before we draw strong conclusions from the findings of this paper.

7 References

Akhtari, M., Bau, N., & Laliberté, J.-W. P. (2020). Affirmative action and pre-college human capital. *National Bureau of Economic Research*.

Arcidiacono, P. (2005). Affirmative action in higher education: How do admission and financial aid rules affect future earnings? *Econometrica, 73*(5), 1477–1524.

Black, S. E., Denning, J. T., & Rothstein, J. (2023). Winners and losers? The effect of gaining and losing access to selective colleges on education and labor market outcomes. *American Economic Journal: Applied Economics, 15*(1), 26–67.

Bleemer, Z. (2022). Affirmative action, mismatch, and economic mobility after California's Proposition 209. *The Quarterly Journal of Economics, 137*(1), 115–160.

Hinrichs, P. (2012). The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities. *Review of Economics and Statistics, 94*(3), 712–722.

Hinrichs, P. (2014). Affirmative action bans and college graduation rates. *Economics of Education Review, 42*, 43–52.

Howell, J. S. (2010). Assessing the impact of eliminating affirmative action in higher education. *Journal of Labor Economics, 28*(1), 113–166.

Khanna, G. (2020). Does affirmative action incentivize schooling? Evidence from India. *Review of Economics and Statistics, 102*(2), 219–233.

Lei de Cotas. (2012). *Lei N 12.711*.

Mello, U. (2022). Centralized admissions, affirmative action, and access of low-income students to higher education. *American Economic Journal: Economic Policy, 14*(3), 166–197.

Mello, U. (2023). Affirmative action and the choice of schools. *Journal of Public Economics, 219*, 104824.

OECD. (2020). PISA 2018 results (Volume V): Effective policies, successful schools. *OECD Publishing*. https://doi.org/10.1787/ca768d40-en

Otero, S., Barahona, N., & Dobbin, C. (2021). Affirmative action in centralized college admission systems: Evidence from Brazil. *Unpublished manuscript*.

8 Figures

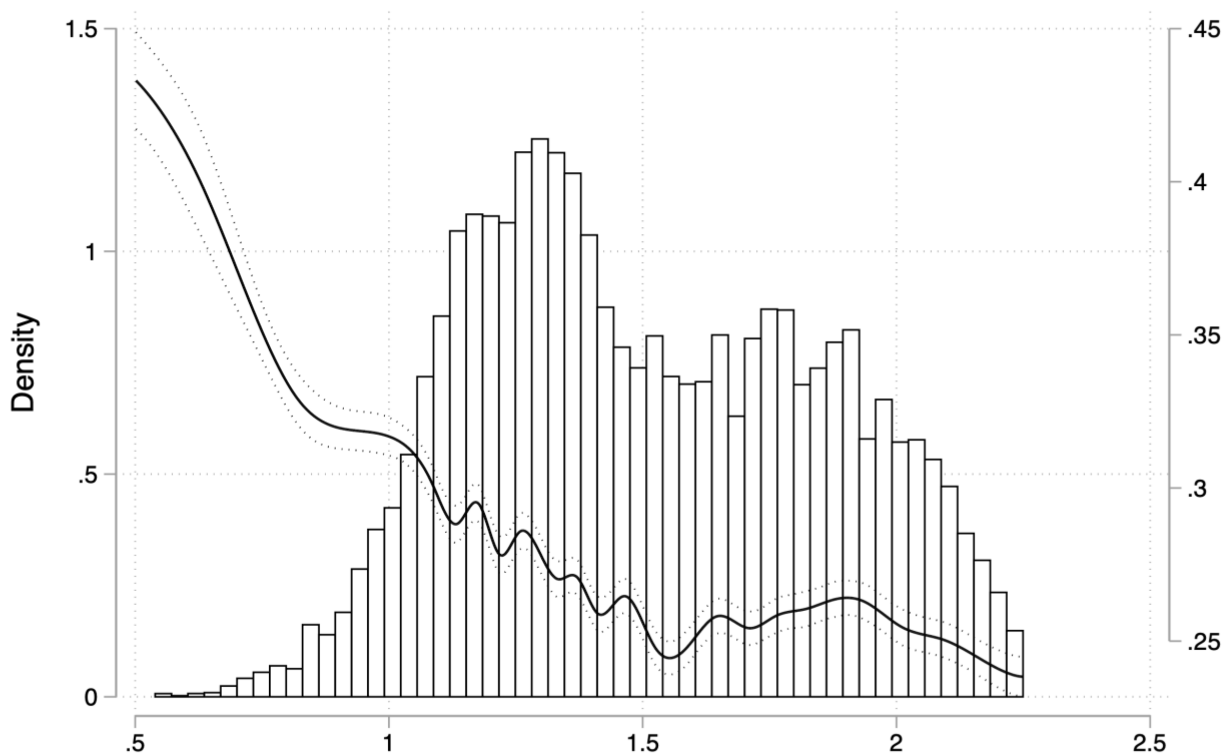

Figure 1. Relationship between z_i, the sixth root of the distance to the nearest college that offers at least 200 seats (horizontal axis) and y_i, college enrollment (right vertical axis). Density of z_i is also plotted with the value of the density on the left vertical axis.

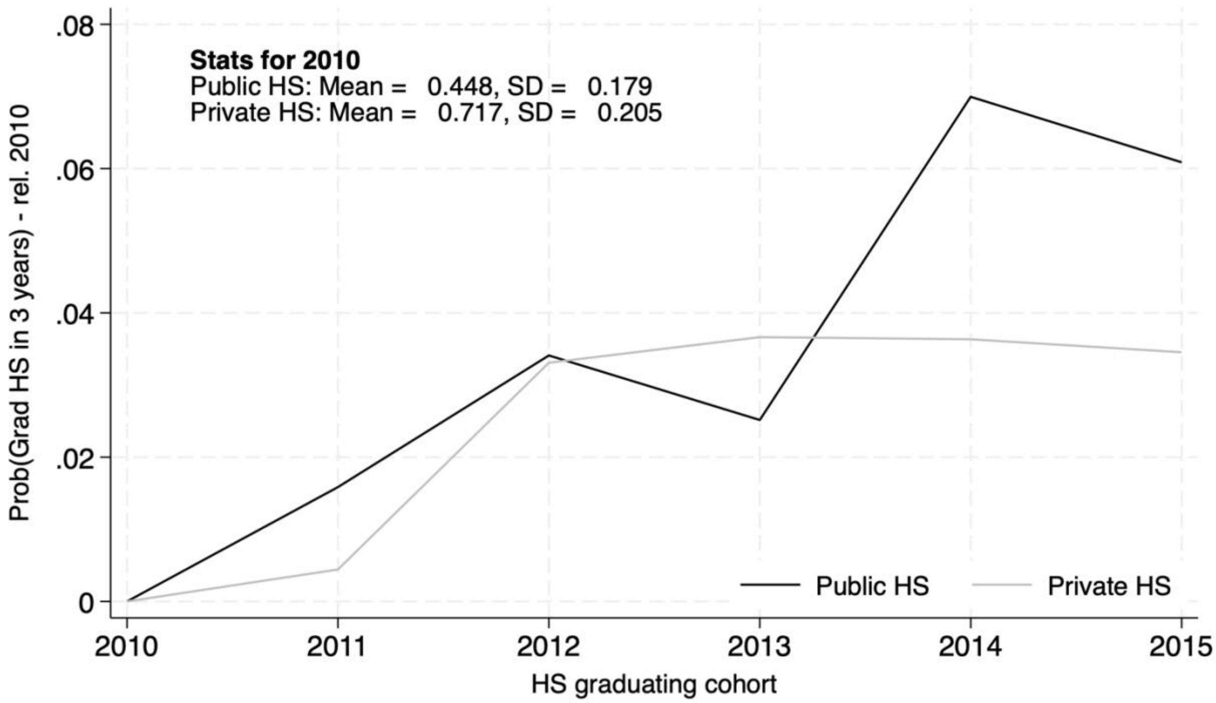

Figure 1. Probability of graduating high school on time by cohort and type of high school. Graphs are shifter vertically so that the mean for each school type in 2010 is zero.

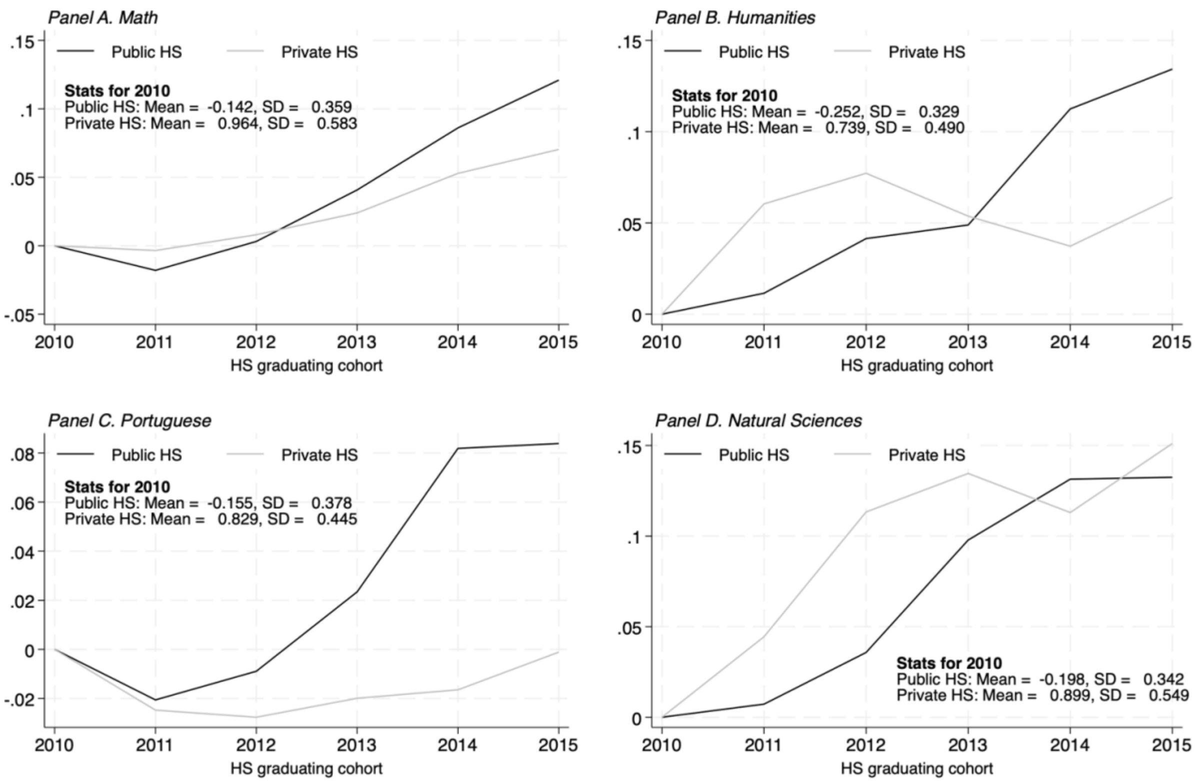

Figure 2. Average test score for top decile of students within school-cohort cell. Graphs are shifted vertically so that the mean of the variable for each type of school in 2010 is mapped to zero.

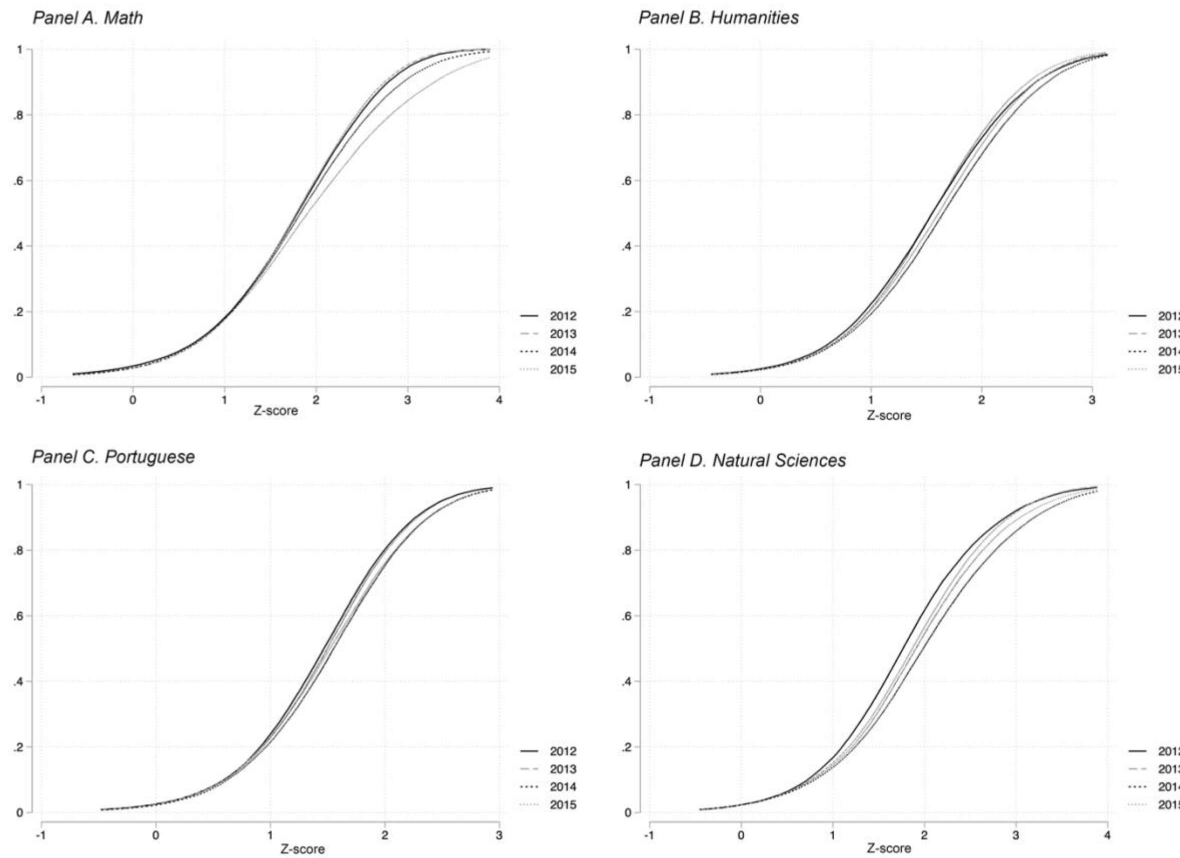

Figure 3. Empirical CDF of test scores of private-high school students who enrolled in a public college, by year of enrollment and by subject.

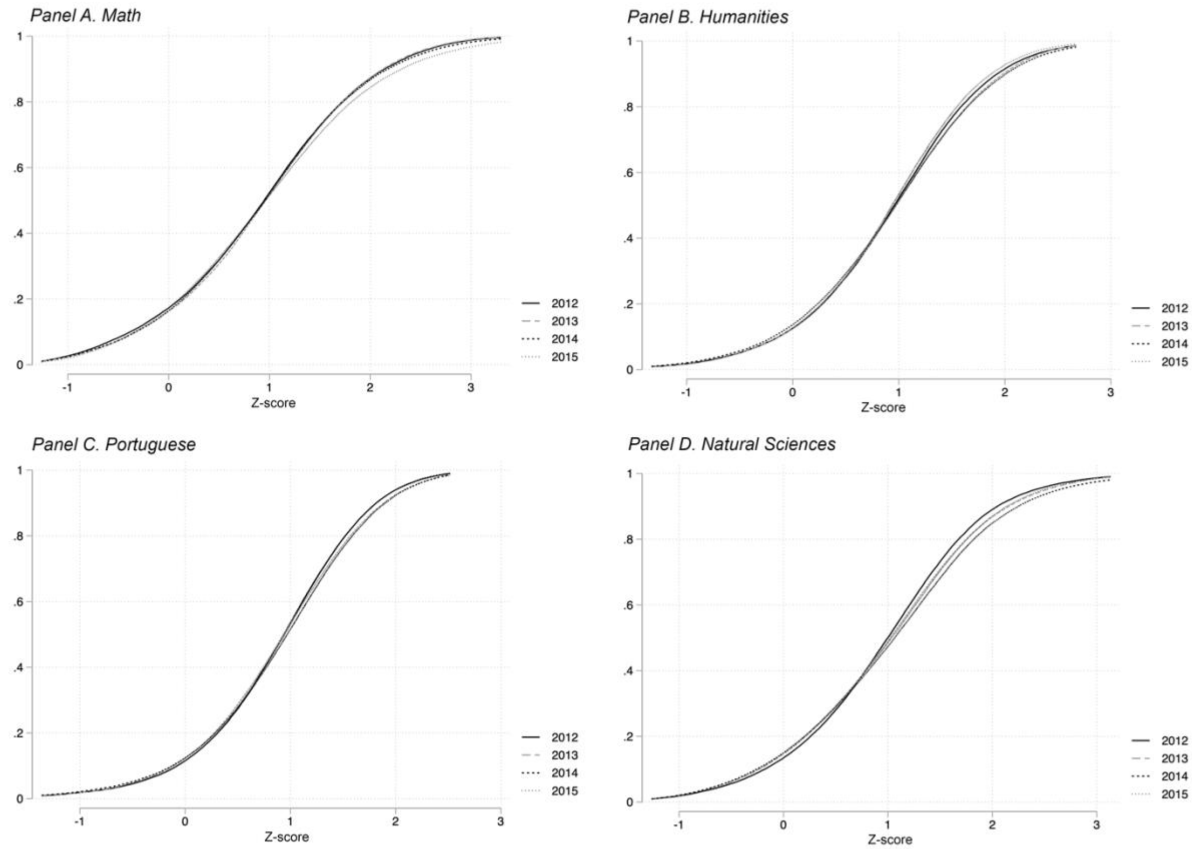

Figure 4. Empirical CDF of test scores of public-high school students who enrolled in a public college, by year of enrollment and by subject.

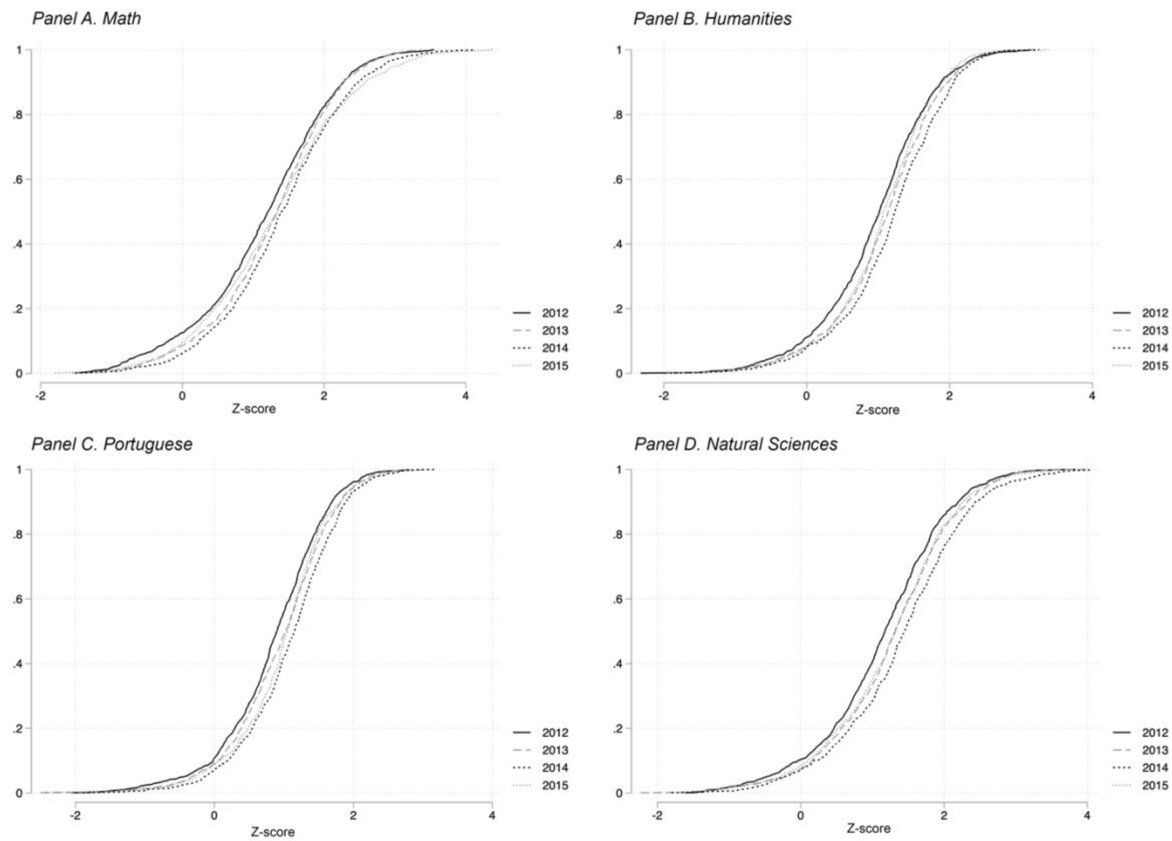

Figure 5. Empirical CDF of test scores of public high school students who enrolled in a public college that was not affected by the Quotas Law, by year of enrollment and by subject.

9 Tables

Table 1. Descriptive Statistics

Variable	Public Schools (1)	Private Schools (2)	Total (3)
A. Demographics and HS Graduation			
N Students	2,132,671	297,147	2,429,818
Racial Minority	0.231	0.092	0.214
Age on year of HS enrollment	17.361	15.433	17.125
Graduated HS in 3 years	0.448	0.717	0.481
Graduated HS in 4 years	0.513	0.775	0.545
Graduated HS in 6 years	0.533	0.782	0.563
B. Measured within 3 years of HS enrollment			
Enroll in college	0.077	0.386	0.115
Enroll in private college	0.057	0.258	0.082
Enroll in public college	0.021	0.142	0.036
Enroll in STEM program	0.014	0.083	0.023
Enroll in STEM program in public college	0.005	0.042	0.010
Enroll in Teaching program	0.013	0.028	0.015
Took ENEM at least once	0.345	0.619	0.378
Avg. Humanities score (top 10%)	-0.252	0.739	-0.131
Avg. Natural Sciences score (top 10%)	-0.198	0.899	-0.064
Avg. Math score (top 10%)	-0.142	0.964	-0.006
Avg. Language score (top 10%)	-0.155	0.829	-0.035
C. Measured within 6 years of HS enrollment			
Enroll in college	0.166	0.620	0.221
Enroll in private college	0.132	0.437	0.169
Enroll in public college	0.041	0.233	0.065
Enroll in STEM program	0.031	0.140	0.045
Enroll in STEM program in public college	0.010	0.070	0.017
Took ENEM at least once	0.394	0.649	0.425
Avg. Humanities score (top 10%)	-0.173	0.808	-0.053
Avg. Natural Sciences score (top 10%)	-0.123	0.982	0.012
Avg. Math score (top 10%)	-0.087	1.013	0.048
Avg. Language score (top 10%)	-0.100	0.881	0.020

Table 2. Timing of High School Graduation and College Enrollment

Event	Did not Happen (1)	2011 (2)	2012 (3)	2013 (4)	2014 (5)	2015 or after (6)
Graduate HS	1,261,753	1,335,194	179,777	41,058	13,752	22,676
	[44.20]	[46.78]	[6.30]	[1.44]	[0.48]	[.79]
		{83.84}	{11.29}	{2.58}	{0.86}	{1.42}
Enroll in College	2,029,108	310,429	179,784	106,714	73,344	154,831
	[71.09]	[10.88]	[6.30]	[3.74]	[2.57]	[5.42]
		{37.62}	{21.79}	{12.93}	{8.89}	{18.77}
Enroll in Public College	2,611,561	100,395	52,605	27,025	16,446	46,178
	[91.50]	[3.52]	[1.84]	[0.95]	[0.58]	[1.62]
		{41.37}	{21.68}	{11.14}	{6.78}	{19.03}

Notes. Table reports the distribution of timing of high school graduation, enrollment in college and enrollment in a public college for the 2010 graduating cohort. Column 1 reports the number of students for which the corresponding event is not observed in the period for which I have data. Columns 2, 3, 4, 5 and 6 report the number of students for which the corresponding event happened in 2011, 2012, 2013, 2014 and 2015 or after, respectively. The percentage of the cohort to which each number number corresponds is reported in brackets. The number in braces is the percentage of the sample of students for which the event is observed – in other words, it reports the distribution conditional on the event being observed.

Table 3. High School Graduation of Private School Students

	Graduate in 3 years (1)	Graduate in 6 years (2)
Panel A. Pre-Policy		
ATT_{2011}	0.011	0.010
	(0.050)	(0.049)
Panel B. Post-Policy		
ATT_{2012}	0.00473	0.00545
	(0.061)	(0.063)
ATT_{2013}	–0.008	–0.002
	(0.048)	(0.047)
ATT_{2014}	–0.013	–0.014
	(0.049)	(0.047)
ATT_{2015}	–0.002	–0.006
	(0.054)	(0.054)
Observations	55,604	55,604
Mean (2010)	0.686	0.751

Notes. Standard errors are reported in parenthesis. The outcome in column 1 is an indicator for graduating high school on time, i.e., in 3 years and the outcome in column 2 is an indicator for graduating high school in 6 years. The pre-policy estimate uses 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students.

Table 4. College Enrollment of Private School Students

	Any College (1)	Private College (2)	Public College (3)	STEM Program (4)
Panel A. Pre-Policy				
ATT_{2011}	–0.017	–0.023	–0.010	0.000
	(0.040)	(0.034)	(0.038)	(0.020)
Panel B. Post-Policy				
ATT_{2012}	–0.012	–0.018	0.011	–0.005
	(0.036)	(0.032)	(0.037)	(0.020)
ATT_{2013}	–0.022	–0.004	–0.006	–0.008
	(0.035)	(0.031)	(0.033)	(0.018)
ATT_{2014}	0.014	0.036	–0.022	–0.001
	(0.035)	(0.030)	(0.035)	(0.018)
ATT_{2015}	0.006	0.023	–0.015	0.002
	(0.037)	(0.028)	(0.037)	(0.017)
Observations	55,604	55,604	55,604	55,604
Mean (2010)	0.596	0.423	0.216	0.132

Notes. Standard errors are reported in parenthesis. The outcomes in columns 1, 2, 3 and 4 are indicators for enrolling in college/university, enrolling in a private college/university, enrolling in a public college/university and enrolling in a STEM program. All outcomes are measured up to 5 years after enrollment in high school, e.g. if the student enrolled in college 7 years after enrolling in high school, then the outcome in column 1 will be zero. The pre-policy estimates use 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students.

Table 5. Test-taking Behavior and Exam Scores of Private-School Students

	Took ENEM (1)	Math (2)	Portuguese (3)	Humanities (4)	Science (5)
Panel A. Pre-Policy					
ATT_{2011}	–0.019	–0.091	–0.047	–0.077	–0.071
	(0.0407)	(0.138)	(0.101)	(0.116)	(0.138)
Panel B. Post-Policy					
ATT_{2012}	0.014	0.035	–0.020	–0.018	–0.001
	(0.036)	(0.137)	(0.105)	(0.118)	(0.146)
ATT_{2013}	–0.024	–0.002	–0.045	–0.036	–0.025
	(0.033)	(0.136)	(0.095)	(0.107)	(0.136)
ATT_{2014}	–0.008	0.002	–0.028	–0.019	–0.041
	(0.034)	(0.145)	(0.099)	(0.112)	(0.140)
ATT_{2015}	–0.018	–0.003	–0.079	–0.050	–0.063
	(0.033)	(0.149)	(0.101)	(0.113)	(0.146)
Observations	55,604	53,511	53,536	53,554	53,531
Mean (2010)	0.593	0.961	0.848	0.742	0.890
SD (2010)	0.170	0.610	0.454	0.510	0.572

Notes. Standard errors are reported in parenthesis. The outcome in column 1 is an indicator for taking ENEM. Test scores are measured as the average for the top 10% of students within a school and cohort. The outcomes in columns 2, 3, 4 and 5 are the top 10% average scores in math, portuguese, humanities and science respectively. All outcomes are measured up to 3 years after enrollment in high school, so scores on tests taken after that are not counted. The pre-policy estimates use 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students.

Table 6. High School Graduation of Public School Students

	Graduate in 3 years (1)	Graduate in 6 years (2)
Panel A. Pre-policy		
ATT_{2011}	0.001	0.007
	(0.015)	(0.015)
Panel B. Post-policy		
ATT_{2012}	0.0160	0.009
	(0.015)	(0.014)
ATT_{2013}	0.032	0.021
	(0.014)	(0.014)
ATT_{2014}	0.023	0.010
	(0.014)	(0.013)
ATT_{2015}	0.035	0.019
	(0.015)	(0.014)
Observations	126,828	126,828
Mean (2010)	0.401	0.488

Notes. Standard errors are reported in parenthesis. The outcome in column 1 is an indicator for graduating high school on time, i.e., in 3 years and the outcome in column 2 is an indicator for graduating high school in 6 years. The pre-policy estimate uses 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students.

Table 7. College Enrollment of Public School Students

	Any College (1)	Private College (2)	Public College (3)	STEM Program (4)
Panel A. Pre-Policy				
ATT_{2011}	–0.003	–0.005	–0.001	0.001
	(0.012)	(0.009)	(0.006)	(0.004)
Panel B. Post-Policy				
ATT_{2012}	0.004	0.001	0.004	–0.003
	(0.013)	(0.010)	(0.007)	(0.004)
ATT_{2013}	0.007	0.005	0.001	0.000
	(0.012)	(0.010)	(0.006)	(0.004)
ATT_{2014}	0.009	0.008	0.003	0.001
	(0.012)	(0.009)	(0.007)	(0.004)
ATT_{2015}	0.011	0.009	0.005	0.004
	(0.012)	(0.009)	(0.007)	(0.004)
Observations	126,828	126,828	126,828	126,828
Mean (2010)	0.167	0.136	0.037	0.032

Notes. Standard errors are reported in parenthesis. The outcomes in columns 1, 2, 3 and 4 are indicators for enrolling in college/university, enrolling in a private college/university, enrolling in a public college/university and enrolling in a STEM program. All outcomes are measured up to 5 years after enrollment in high school. The pre-policy estimates use 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students.

Table 8. Test-taking Behavior and Exam Scores of Public School Students

	Took ENEM (1)	Math (2)	Portuguese (3)	Humanities (4)	Science (5)
Panel A. Pre-Policy					
ATT_{2011}	0.009	–0.026	–0.034	–0.032	–0.012
	(0.015)	(0.032)	(0.034)	(0.033)	(0.031)
Panel B. Post-Policy					
ATT_{2012}	0.009	0.019	0.047	0.035	0.014
	(0.016)	(0.031)	(0.030)	(0.031)	(0.030)
ATT_{2013}	0.004	0.035	0.025	0.029	0.006
	(0.015)	(0.029)	(0.030)	(0.029)	(0.028)
ATT_{2014}	0.001	0.023	0.028	0.025	0.017
	(0.014)	(0.030)	(0.030)	(0.029)	(0.029)
ATT_{2015}	0.004	0.059	0.035	0.030	0.023
	(0.015)	(0.029)	(0.029)	(0.029)	(0.028)
Observations	126,828	115,597	115,794	116,162	115,969
Mean (2010)	0.326	-0.122	-0.075	-0.219	-0.186
SD (2010)	0.148	0.368	0.364	0.339	0.353

Notes. Standard errors are reported in parenthesis. The outcome in column 1 is an indicator for taking ENEM. Test scores are measured as the average for the top 10% of students within a school and cohort. The outcomes in columns 2, 3, 4 and 5 are the top 10% average scores in math, portuguese, humanities and science respectively. All outcomes are measured up to 3 years after enrollment in high school, so scores on tests taken after that are not counted. The pre-policy estimates use 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students.

A. Changes in the Supply of Public College

One potential threat to the parallel trends assumption occurs if changes in the number of college seats available in the vicinity of high schools is heterogeneous across exposed and non-exposed regions. To check whether this is a concern, I first measure the number of public college seats available within a radius of $30km$ around a school s in 2011 and 2015, which I denote by u_s^{2011} and u_s^{2015}. I then compute the values $y_s = (u_s^{2015})^{1/6} - (u_s^{2011})^{1/6}$, where I use the sixth root as an approximation to the log transformation because some schools do not have any public college seats within $30km$ in 2011 or in 2015. Finally, I estimate the nonparametric regression model given by

$$y_s = \sum_{j=1}^{J} f_j(x_{sj}) + \delta_{p(s)} + \varepsilon_s.$$

In this model, $\delta_{p(s)}$ is a state fixed effect and for each $j = 1, \dots, J$, the variable x_{sj} is a characteristic of school s—specifically, I include the average income and population density of the census tracts where the school is located, the exposure variable defined in section 4.3 and the number of high school students within a $30km$ radius. Additive separability between the school characteristics and the fixed effects is assumed for tractability, and each $f_j(\cdot)$ is estimated nonparametrically using a series regression estimator with a third-degree spline basis. Figure A1 reports the estimates of each of the $f_j(\cdot)$—note that because of the fixed effects the intercepts are not identified and only the slopes of the lines plotted in the figure are of interest. We observe that net of fixed effects, the growth of the supply of public college seats are mostly explained by the size of the potential market for college, which is proxied by the number of high school students within a $30km$ radius of the school. On the other hand, the exposure variable almost none of the outcome. Thus, there is little evidence that this could be a confounding factor.

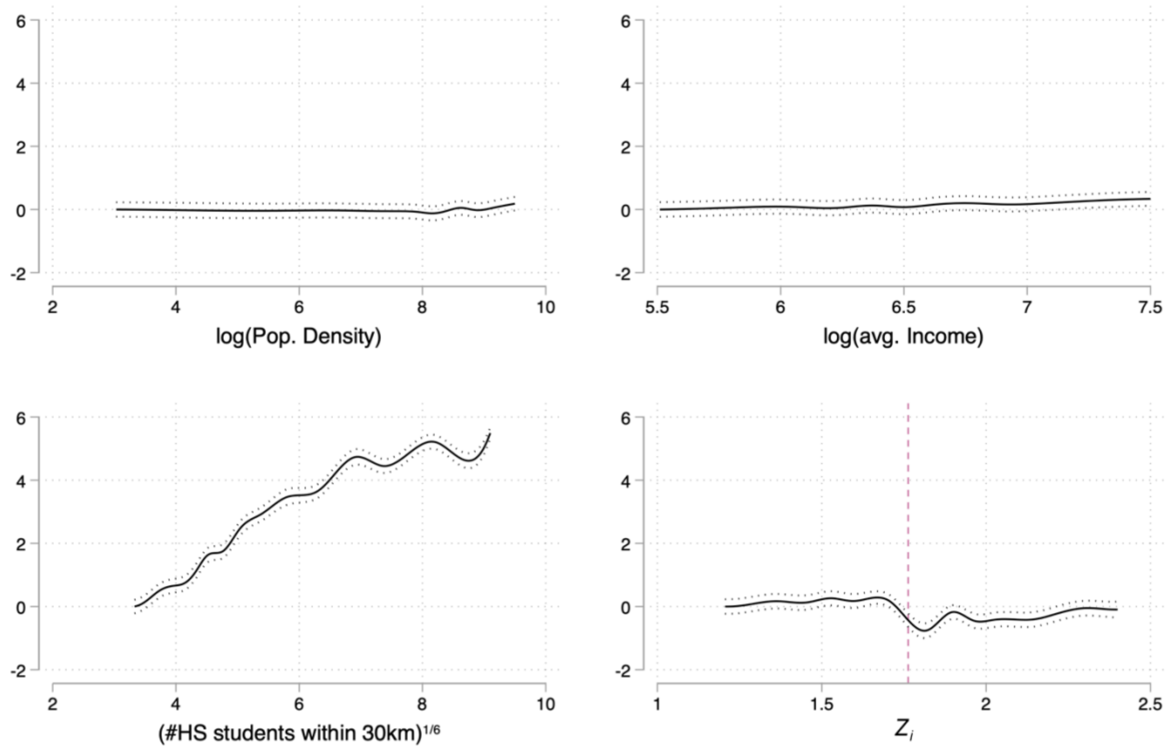

Figure A1. Nonparametric estimates of components of conditional mean function. Dotted lines correspond to pointwise confidence intervals constructed using Eicker-Huber-White standard errors. Because the confidence intervals do not account for misspecification, and assume undersmoothing, they are likely to underestimate the uncertainty of the estimates.

B. Robustness Checks

In this appendix, I report the estimated effects on high school graduation and college enrollment outcomes using only state fixed effects and the number of high school students within a $30km$ radius as the conditioning variables. Tables B1 and B2 report the results for high school graduation and college outcomes of private high school students, respectively. Tables B3 and B4 report the results for public school students' high school graduation and college enrollment outcomes, respectively.

Table B1. High School Graduation – Private High Schools

	Graduate in 3 years (1)	Graduate in 6 years (2)
Panel A. Pre-Policy		
ATT_{2011}	0.00335	0.00295
	(0.0479)	(0.0468)
Panel B. Post-Policy		
ATT_{2012}	0.00692	0.0118
	(0.0589)	(0.0593)
ATT_{2013}	–0.000702	0.000814
	(0.0463)	(0.0454)
ATT_{2014}	–0.00390	–0.00777
	(0.0465)	(0.0452)
ATT_{2015}	0.00990	0.00523
	(0.0513)	(0.0509)
Observations	55,604	55,604
Mean (2010)	0.686	0.751

Notes. Standard errors are reported in parenthesis. The outcome in column 1 is an indicator for graduating high school on time, i.e., in 3 years and the outcome in column 2 is an indicator for graduating high school in 6 years. The pre-policy estimate uses 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students. Conditioning variables are state fixed-effects, which are interacted with the sixth root of the number of students within a 30km radius of each high school.

Table B2. College Enrollment – Private High Schools

	Any College (1)	Private College (2)	Public College (3)	STEM Program (4)
Panel A. Pre-Policy				
ATT_{2011}	–0.0260	–0.0237	–0.0230	–0.00558
	(0.0423)	(0.0350)	(0.0447)	(0.0214)
Panel B. Post-Policy				
ATT_{2012}	–0.0127	–0.0170	0.00779	–0.00831
	(0.0384)	(0.0337)	(0.0437)	(0.0203)
ATT_{2013}	–0.00919	–0.00395	0.00793	–0.00578
	(0.0370)	(0.0319)	(0.0382)	(0.0192)
ATT_{2014}	0.0171	0.0343	–0.0150	0.00112
	(0.0372)	(0.0300)	(0.0405)	(0.0189)
ATT_{2015}	0.0171	0.0237	–0.00322	0.00498
	(0.0385)	(0.0295)	(0.0411)	(0.0182)
Observations	55,604	55,604	55,604	55,604
Mean (2010)	0.596	0.423	0.216	0.132

Notes. Standard errors are reported in parenthesis. The outcomes in columns 1, 2, 3 and 4 are indicators for enrolling in college/university, enrolling in a private college/university, enrolling in a public college/university and enrolling in a STEM program. All outcomes are measured up to 5 years after enrollment in high school, e.g. if the student enrolled in college 7 years after enrolling in high school, then the outcome in column 1 will be zero. The pre-policy estimates use 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students. Conditioning variables are state fixed-effects, which are interacted with the sixth root of the number of students within a 30km radius of each high school.

Table B3. High School Graduation – Public High Schools

	Graduate in 3 years (1)	Graduate in 6 years (2)
Panel A. Pre-policy		
ATT_{2011}	0.00447	0.00870
	(0.0150)	(0.0145)
Panel B. Post-policy		
ATT_{2012}	0.0135	0.00836
	(0.0142)	(0.0139)
ATT_{2013}	0.0310	0.0215
	(0.0143)	(0.0138)
ATT_{2014}	0.0231	0.0117
	(0.0136)	(0.0129)
ATT_{2015}	0.0328	0.0187
	(0.0147)	(0.0143)
Observations	126,828	126,828
Mean (2010)	0.401	0.488

Notes. Standard errors are reported in parenthesis. The outcome in column 1 is an indicator for graduating high school on time, i.e., in 3 years and the outcome in column 2 is an indicator for graduating high school in 6 years. The pre-policy estimate uses 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students. Conditioning variables are state fixed-effects, which are interacted with the sixth root of the number of students within a 30km radius of each high school.

Table B4. College Enrollment – Public High Schools

	Any College (1)	Private College (2)	Public College (3)	STEM Program (4)
Panel A. Pre-Policy				
ATT_{2011}	–0.00343	–0.00478	–0.00123	0.000784
	(0.0137)	(0.0106)	(0.00676)	(0.00421)
Panel B. Post-Policy				
ATT_{2012}	0.00368	0.000523	0.00411	–0.00319
	(0.0146)	(0.0114)	(0.00721)	(0.00459)
ATT_{2013}	0.00763	0.00603	0.00149	2.30e-05
	(0.0136)	(0.0107)	(0.00666)	(0.00435)
ATT_{2014}	0.0103	0.00875	0.00312	0.00234
	(0.0138)	(0.0106)	(0.00692)	(0.00415)
ATT_{2015}	0.0134	0.0108	0.00450	0.00429
	(0.0135)	(0.0105)	(0.00695)	(0.00401)
Observations	126,828	126,828	126,828	126,828
Mean (2010)	0.167	0.136	0.037	0.032

Notes. Standard errors are reported in parenthesis. The outcomes in columns 1, 2, 3 and 4 are indicators for enrolling in college/university, enrolling in a private college/university, enrolling in a public college/university and enrolling in a STEM program. All outcomes are measured up to 5 years after enrollment in high school. The pre-policy estimates use 2010 as the control cohort and 2011 as the treated cohort. The reported mean corresponds to the mean of the outcome for the treated group in 2010. Number of observations correspond to the number of school-cohort cells used in estimation. Each school-cohort cell is weighted by the number of students. Conditioning variables are state fixed-effects, which are interacted with the sixth root of the number of students within a 30km radius of each high school.