College Expansion and Unequal Access to Education in Peru

José Flor-Toro Matteo Magnaricotte *
Northwestern University

October 2021

Click here for the most recent version

Abstract

Enrollment gaps are pervasive in developing countries, despite public investment and legislation aimed at democratizing access to college. We study the effects of opening new college campuses in underserved areas, a commonly proposed policy to reduce such gaps. Using Peruvian census data to estimate a difference in differences model, we find that enrollment increased by about $1 p.p.$ or 10% in the short term. However, estimated effects for minority students are only half the size of others, widening preexisting gaps. To understand the drivers of this result and simulate counterfactuals, we assemble a new administrative dataset on college applications and build a model of education demand with heterogeneity in preferences and probability of admission. The results show that initial advantage and meritocratic criteria interact to reinforce educational inequality: even though proximity is highly valued by less-advantaged students, meritocratic admission criteria hinder poor and minority students, who disproportionately attend lower-quality high schools. Our counterfactuals show that addressing high school quality disparities is more likely to reduce college enrollment inequality than further supply expansions.

*We thank Matt Notowidigdo, Chris Udry, Nicola Bianchi, and Gaston Illanes for guidance and support. We thank Fabiola Alba Vivar, Matteo Camboni, Leander Heldring, Tai Lam, Joris Mueller, Michael Porcellacchia, Robert Porter, Mar Reguant, Lorenzo Stanca, Miguel Talamas, Francesca Truffa, for useful discussions and suggestions. Ricardo Sanchez and Mario Gonzales provided help with data and legislative context. Lucia Gomez Llactahuamani provided excellent research assistance.

1 Introduction

From a development perspective, increasing college completion rates is important for regional growth (Barro (1991), and Gennaioli et al. (2013)) and for the creation of upper tail human capital (Mokyr (2005), and Squicciarini and Voigtländer (2015)). At the same time, as recognized by the United Nations in their Sustainable Development Goals (Desa and others (2016)) and in academic research (Chetty et al. (2020)), a college degree is an important factor for economic mobility. To reduce the advantage that students with more initial resources (e.g. family wealth, and more educated or connected parents) have over otherwise similarly talented students, governments often intervene to increase equality of access with policies like affirmative action (Arcidiacono and Lovenheim (2016)), top-percent policies[1] (Bleemer et al. (2020)), and subsidized tuition for targeted groups.[2]

This problem is particularly acute in developing countries, which often have limited ability to implement targeted policies, smaller budgets, and very stratified societies: despite their enrollment rates nearing those of richer nations, many developing countries are still facing large gaps in educational attainment. Data on 160 developing countries included the WIDE database[3] show that their gaps in enrollment rates between the richest and poorest quintiles are on average 1.5 times the size of their mean enrollment rate. A common policy aimed at reducing gaps while increasing overall enrollment is the creation of new public colleges in areas with limited higher education alternatives.[4] This policy has not received much scholarly attention and has ambiguous implications for equality of access: even though less-advantaged students are more likely to value proximity of college options, the meritocratic allocation of scarce seats might favor more-advantaged students with access to better preparation.[5]

We study the consequences of a rapid expansion of the higher education sector in Peru,

[1] Top percent policies guarantee admission at selective colleges to applicants whose grades ranked at the top of their high school class (e.g. in Texas, in the top 10%).

[2] Hsieh et al. (2019) also makes the case that enabling individuals to pursue their comparative advantage can have large effects on aggregate productivity through an improved allocation of talent.

[3] The World Inequality Database on Education, developed by UNESCO, sources data on education from DHS, MICS, and other household surveys. It can be accessed at https://www.education-inequalities.org/.

[4] See, for example, recent debates in Chile https://opinion.cooperativa.cl/opinion/educacion/educacion-expectativas-del-nuevo-gobierno/2021-10-05/130743.html and Argentina https://www.pagina12.com.ar/359354-crear-dos-tres-muchas-universidades-publicas .

[5] Heckman and Mosso (2014) notes that universal provision of policies need not promote equality of outcomes when complementarities are present and more advantaged households are better able to take advantage of them.

whereby new campuses were opened in areas with no prior private or public university option available. Peru, similarly to most Latin American countries[6], is currently facing large attainment gaps related to ethnicity, gender, and socio-economic status (SES).[7] We quantify the effects that the creation of 69 new college campuses since 1960 had on enrollment of the local population and of several subgroups. These colleges offered free education but had limited capacity, with only 1 in 5 applicants admitted through an entrance exam.

We explore the impact of college openings in the period 1960-2009 on enrollment and completion, addressing two main empirical challenges. To deal with non-random location of new campuses, we use a difference in differences approach, comparing changes in enrollment rates for different cohorts in treated areas with corresponding changes in non-treated areas. Rather than relying on survey data, with potential representativeness issues, we use census data to obtain precise measurement of enrollment and completion rates. When a new campus opens up, enrollment increases among cohorts 18 or younger[8] by approximately 1 percentage points (p.p.) or 8% in the short run ($2.2 p.p.$ or 17% in the long run). This effect is very heterogeneous: the ethnic majority and women have higher increases in enrollment and completion. In particular, we find an increase in enrollment among ethnic minorities of just 0.6p.p., less than half that of the majority.

Opening of new campuses affects outcomes through a decrease in distance from the closest option and through changes in the probability of admission. To tease apart these two channels, we develop a discrete choice model for educational demand that incorporates the admission process of public colleges. We estimate it using a data that tracks individuals in the cohort graduating from high school in 2018 throughout their educational career, including college application, admission results, and enrollment.[9]

We estimate a flexible specification of indirect utility where distance affects preferences heterogeneously. This allows for groups of different backgrounds to respond more or less

[6] Ferreyra et al. (2017) provides a description of the sector for Peru and other LATAM countries.

[7] 55% of Latin American and Caribbean countries have inclusive education as a priority in their education sector plans or strategies (UNESCO (2020)) and the Peruvian government has recently recognized the problem by including inclusive education as an objective in the "National plan for higher, technical, and productive education" (Ministerio de Educación (2020)). See also SUNEDU (2020) and Sánchez et al. (2021) for reports on stratification of access in Peru.

[8] High school graduation is at 17 in Peru. More details about the educational system can be found in Section 2.

[9] While census data allows precise measurement and spans multiple cohorts, it does not include questions regarding college applications.

to changes in distance from college. Similarly, admission probabilities vary depending on students' characteristics and on proximity to college. As a first counterfactual, we simulate the opening of new public campuses in areas that do not have any. This simulation serves two purposes: first, it lets us benchmark our model by comparing the predicted effects to those estimated from the difference in differences model; second, it provides a way to learn about the discussed mechanisms. We find that the effects predicted by the demand model are comparable to those found in the previous section, with the same heterogeneity pattern for the ethnicity and gender dimensions. Reflecting such heterogeneity, we find that opening college campuses *increases* ethnicity and gender attainment gaps by 9% and 6% respectively, while it leaves unaffected the socioeconomic status (SES) gap. We show that proximity to campus increases applications more for individuals with low-SES, but this gap-reducing effect is offset by the admission process, which favors more advantaged students.

We also find that college openings increase the ethnic gap *both* through application decisions and probability of admission conditional on application. We hypothesize that this might be due to differences in prior preparation that make college education less appealing to minority students and reduce their admission chances. We show that minority students are over-represented at low-quality high schools and simulate a policy that reduces quality gaps at the secondary level: this closes half of the ethnic gap, showing the promise of policies that address upstream inequality. Finally, we show that the admission process increases gaps along all measured dimensions in both simulated counterfactuals. This is explained by complementarities between the proposed policies and characteristics of the advantaged groups: e.g. applicants whose parents are more educated will benefit more than others from high-quality education, making them more likely to be admitted to college.[10]

Our analysis is connected to several strands of literature. Most directly, it relates to studies estimating the educational effects of college openings. Russell et al. (2021) studies long-run effects of openings in the US by comparing sites where a college campus was open between 1839 and 1954 with runner-up locations, finding that cumulative exposure to colleges significantly affects college attainment today. Jagnani and Khanna (2020) studies the effects of the establishment of elite public colleges in India on high school completion: while no

[10] Heckman and Mosso (2014) considers complementarities between parental education and public investment to be a hypothesis in need of further validation, reporting Pop-Eleches and Urquiola (2013) and Gelber and Isen (2013) as examples of studies contributing evidence of its plausibility.

effect on college enrollment is found due to the extreme selectivity of these colleges, the concurrent public investment in electricity, roads, and water services led to increased attainment among school-aged kids. Oppedisano (2011) uses an IV approach to study the Italian expansion of college supply in the period 1995–98, finding positive enrollment effects but a decline in performance. [11,12]

We contribute to this strand of literature along several dimensions. First, we assemble a dataset with 69 college openings staggered over time and use a difference in differences approach to deal with non-random location. We focus on openings in areas with no previously-available option. These two features and the precision in the timing of opening set our study apart from the existing literature. Second, census data allows us to measure enrollment and completion rates with precision at the province-cohort level, thus overcoming typical limitations of survey or aggregate data.[13] Third, we focus on heterogeneous effects of the policy and its implications for inclusive education.[14] Additionally, our setting is particularly relevant for policymakers in developing countries.

Second, we provide analysis of an understudied policy tool, often used with the goal of reducing attainment gaps of URM and lower-SES students. In the US context, the two policies for equitable access to higher education that have been studied the most are affirmative action (see Arcidiacono and Lovenheim (2016) for a recent review) and "top percent" programs (Black et al. (2020) and Bleemer et al. (2020) are recent examples). In Brazil, Mello (2019) studies the effects and interaction of affirmative action and centralization of admission. Other papers studying interventions to reduce attainment gaps have shown heterogeneous effects (e.g. Carrell and Sacerdote (2017)) and the potential for the emergence of private responses that offset the equity gains from the policy (e.g. Chatterjee et al. (2020)).

We contribute by documenting the effects of an expansion of supply in higher education

[11] In another study of the Italian college system, Bianchi (2020) uses a different kind of expansion in college access, leveraging a relaxation of admission requirements for less advantaged students to estimate the impact on inframarginal students' learning.

[12] A number of papers, following the seminal Card (1993), used exposure to college as an instrument to identify the effects of education on several other outcomes. In a famous study, Currie and Moretti (2003) focuses on intergenerational transmission of human capital in the US; more recently, Kyui (2016) and Belskaya et al. (2020) estimate returns to college in Russia.

[13] Jagnani and Khanna (2020) and Oppedisano (2011) use survey data; Russell et al. (2021) uses data aggregated at the level of institutions (IPEDS) or county.

[14] Oppedisano (2011) represents an exception, finding enrollment effects concentrated among less-advantaged, middle-ability individuals.

in previously underserved areas (i.e. with no public or private university available), a policy tool commonly used in developing countries that few studies have analyzed, and focusing on the factors that favor a more equitable distribution of access. We show that effects of campus opening are heterogeneous depending on students' characteristics and on the type of college (private or public). Through the estimation of a model for educational demand, we also study the drivers of our findings, and build counterfactual scenarios for different policies. Our results highlight the role of prior preparation and meritocratic admission criteria in determining the heterogeneous response to openings.[15]

Third, we build on the literature estimating students' demand for education.[16] We develop a new model that emphasizes the interaction between preferences and admission criteria, with heterogeneous preferences for proximity to college. Our estimation procedure is closest to the two-step Maximum Likelihood procedure used in Hastings et al. (2017) and Abdulkadiroğlu et al. (2020). In our counterfactual simulations, we study the effects of (further) college openings and of a reduction in the dispersion of high school quality.[17]

The rest of the paper is organized as follows: in Section 2 we describe the characteristics of the Peruvian educational system and context that are relevant to our analysis; Section 3 describes the main datasets that are used in our empirical analysis; in Section 4 we estimate a difference in differences model and discuss its results. In Section 5 we build and estimate a discrete choice model to learn about students' preferences, provide evidence for different mechanisms and simulate counterfactual policies; Section 6 concludes the paper.

[15]The effects of meritocratic criteria for selection and allocation when initial resources are unequally distributed is another underexplored topic: Arcidiacono et al. (2021) studies how a non-meritocratic selection rule affects the composition of the student body at Harvard; Akbarpour et al. (2020) represents a recent attempt to model market design under non-market allocation rules for scarce public goods.

[16]For early examples, see Alderman et al. (2001) for school and Fuller et al. (1982) for post-secondary education. For more recent examples of papers applying IO tools to education markets see: Neilson (2013); Hastings et al. (2016); Kapor et al. (2017); and Dinerstein and Smith (2014).

[17]Finally, we add to the vast literature on effects of increased education. Our results align qualitatively with previous findings on reductions in fertility (Osili and Long (2008), Duflo et al. (2015)), and on increases in employment (Beuermann et al. (2018) and Bautista et al. (2020)). Interestingly, our results for employment are large and significant only for public colleges, while private ones have non-significant, economically smaller estimates. This latter pattern might be due to differences in quality of public and private institutions.

2 Institutional Context

2.1 Higher education system in Peru

Peru's primary education starts at the age of 6 and lasts 6 years, while secondary schooling starts at 12 and lasts 5 more years.[18] Within two years from high school graduation, most students apply and enroll in majors whose formal length is usually 5 years (see Figure 12 in the Appendix for the full distribution of time to college). In practice, many students take 6-7 years to graduate.

Public universities only charge small administrative fees, while private universities do not face restrictions on tuition setting: for this reason, and thanks to a perception of higher average quality, public colleges, whose admission procedures are decentralized, have much higher ratios of applicants to admitted students than private ones. While no centralized admission system exists, all public colleges admit students mainly through admission tests that are generally held twice a year. Students can take admission tests at multiple colleges and can re-take the test without limits: each attempt requires the payment of a small fee that varies across universities. About 20% of applications at public colleges are successful. Private universities have a greater variance, with only a few of them display a significant degree of selectivity and the rest admitting almost any applicant. Admission procedures at private universities are not standardized. Scholarship availability and financial aid instruments are extremely limited.

The main alternative to colleges are the so-called technical institutes, which correspond loosely to community colleges in the US context: they are in majority private (69% of the total according to the 2020 census of educational institutions) and provide vocational education without the possibility to transfer and obtain a university degree.[19]

Congress is the sole institution endowed with the power to create new public univer-

[18] Peru has an area approximately 3 times the one of California (see Figure 10) and population of roughly 31 millions: it is organized into 25 regions and 196 provinces (approximately corresponding to US states and counties, respectively). Peru is quite representative of several other Latin American countries, with a large share of its population concentrated around the capital and long distances between other populated areas. For a comparison of the Peruvian higher education sector with other Latin American countries, see Ferreyra et al. (2017).

[19] The law N° 30512 of Peru regulates these institutions.

sities.[20,21] Recognizing the need for a more diffused supply of higher education, Peru has opened new public universities and university branches in 35 provinces out of the country's 196 since 1990. In addition, with the *law 882 on the promotion of private investment in education* of 1996, the Peruvian Congress allowed for profit institutions in the higher education market.[22] One of the main purposes for the new regulation included the democratization of access (Cuenca (2015)). The same period saw the creation of many new private universities, for a total of 152 new branches. Figure 1 shows the distribution of university campuses in 2017 in each province.[23]

In 2012 Congress passed a moratorium on the creation of new universities to limit the uncontrolled entry of universities of dubious quality and allow for the institution of minimum quality standards.[24] The moratorium limited the creation of public and private universities alike and prohibited the opening of new branches for already established institutions. Starting in 2015 a new public body, named *SUNEDU*[25], was given supervisory duties over the quality of Peruvian universities. This monitoring activity culminated between 2018 and 2020 with the effective closure of about a third of Peruvian universities, all of them private except for one public college.

In the moment of maximum expansion, in 2017, Peru had about 300 different college campuses, with two thirds of them being private (see panel A of Figure 2 for the evolution over time).

2.2 Access to university

College enrollment rates have grown over time, reflecting the general economic development of Peru and increasing from less than 10% for individuals born in 1940 to almost 30%

[20]The last 30 years have seen a tight relationship between Peruvian politics and higher education, with Congress expanding the role of the state at first and implementing the entry moratorium later on, and with private college owners successfully entering politics as Congress members.

[21]Congress is prohibited from making changes to the budget law or increasing public expenditure.

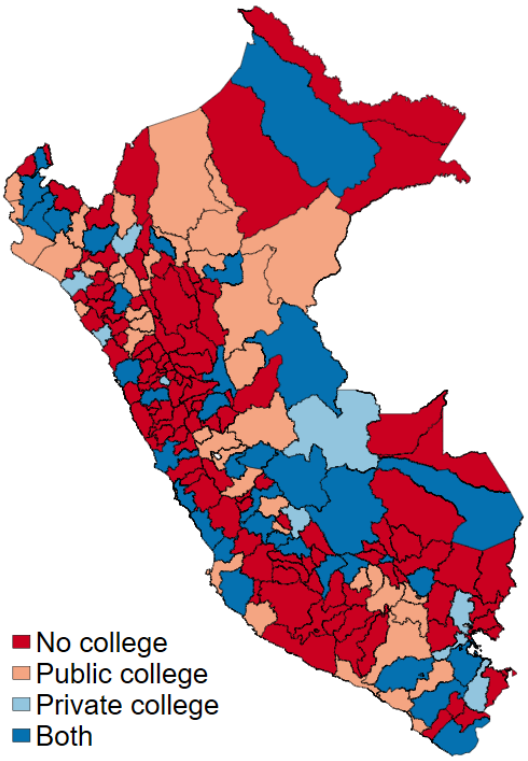
[22]Non profit institutions could be established even before said law, as regulated by article 26 of the same *Ley Universitaria*.

[23]In the empirical analysis of this paper, we will define college campuses as all the buildings belonging to a university in a given province. Consequently, if a university has buildings in two different provinces, we will say that it has two campuses.

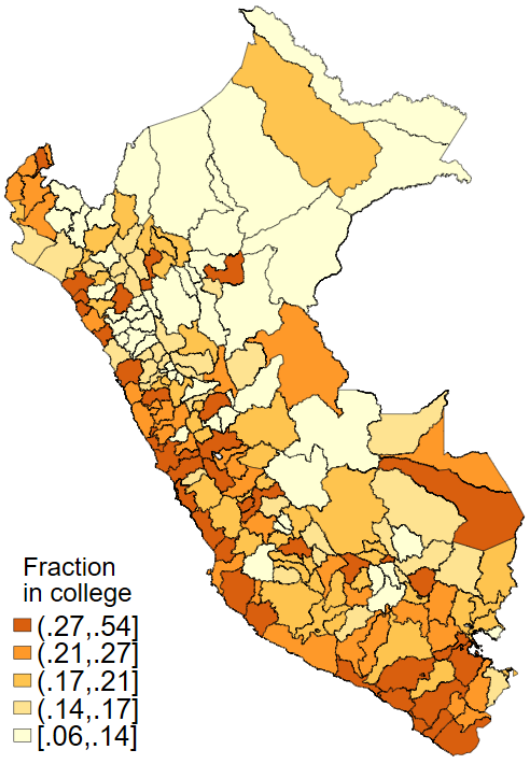
[24]The 2012 moratorium has since been renewed several times. The latest renewal is contained in the Law 31193 of Peru.

[25]Acronym for National Superintendence of Higher Education in Spanish.

Figure 1: Geographical College Availability and Access in Peru



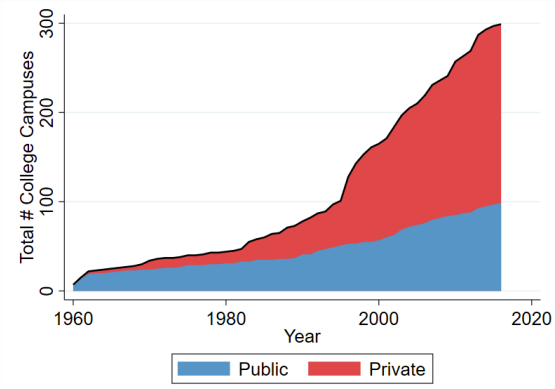
Panel A: Geographical Distribution of College Campuses



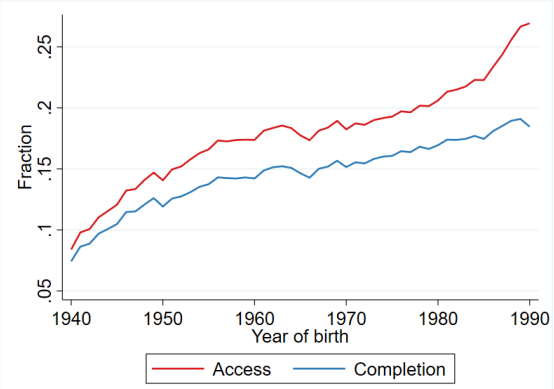
Panel B: Geographical Distribution of College Enrollment

Notes. Panel A shows the geographical distribution of college campuses in the 196 provinces of Peru in 2017. Section 3 reports how the location of each campus of each university is obtained using an administrative census of education institutions. Panel B shows the fraction of individuals born in 1996 declaring college enrollment in the 2017 Census, by province of birth.

Figure 2: College Availability and Access over Time in Peru



Panel A: College Openings over Time, by College Type



Panel B: Access and Completion of College by Cohort of Birth

Notes. Panel A reports the number of college campuses open in each year, by management type (public or private). The process used to obtain the opening year of each campus is described in Section 3. In 1996, Congress authorized the creation of for-profit colleges, leading to a faster increase in college openings. Panel B plots the rates of college enrollment and completion for the cohorts born in the period 1940-1990 in Peru, according to the 2017 Census. Both rates increase over time, with a sharp increase for access for cohorts born after 1980.

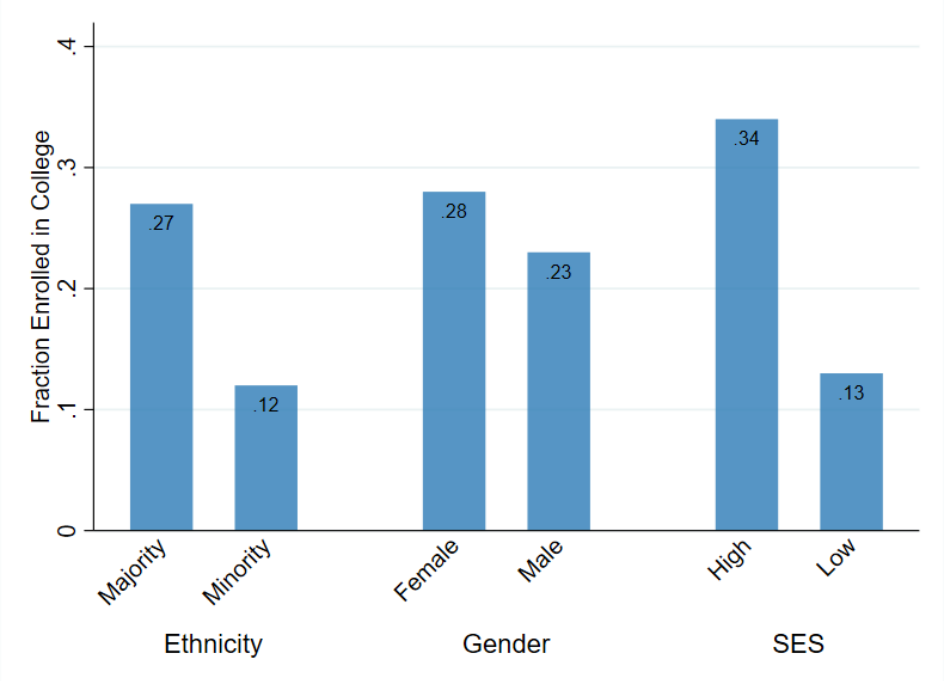
for the 1990 cohort, as shown in panel B of Figure 2. As can be seen in the Figure 2, this growth has been paralleled by the opening of new college campuses. This expansion has allowed for more students to attend colleges close to their places of birth: indeed, student mobility in the higher education setting is fairly limited. Analyses of the 2017 Census of population and the 2010 Census of university students, reported in the Appendix, show that approximately 90% of university students enroll in a college located in their province of birth.

One important political and policy concern has regarded the participation to postsecondary education of ethnic minorities. Peru has a very diverse population, with a quarter of individuals born in the period 1960-2000 self-identifying as belonging to different groups of indigenous populations according to the 2017 Census[26] and indigenous populations have mobilized to increase their access to higher education. This has led to the opening of several "intercultural universities" that provide bilingual education.

As a matter of fact, the racial gap in enrollment is one of the largest, as highlighted in Fig-

[26]Figure 10 in the Appendix shows the percentage of people in each province that learned a language different from Spanish as their first language: we can observe a large level of segregation, with concentration of minorities in the Andean area and clear Spanish-speaking majorities along the coast.

Figure 3: Fraction of Enrollees by Sociodemographic Groups



Notes. Fraction of college enrollees among individuals age 17-24 by sociodemographic groups, according to ENAHO (2014-19).

ure 3 and in other reports (SUNEDU (2020)). Socioeconomic status, represented by parental education, and rurality-urbanicity also emerge as important factors. Females have a smaller but relevant advantage over male students. In 2016 a report of the OECD (OECD (2016)) highlighted gaps in access among the challenges that Peru still faces.[27] Disparities are generally smaller at the primary level and increasing at higher education levels. Panel B of Figure 1 reports college enrollment rates at the province level for the cohort born in 1996, as recorded in the 2017 Census. In response to these gaps, in 2020 the Ministry of Education issued the National Plan for Higher and Technical Education (PNESTP by the Spanish acronym, Ministerio de Educación (2020)) highlighting as the first policy goal the reduction of gaps in access to higher education.[28]

[27] For more descriptive evidence on the state of Peruvian higher education in recent years, see SUNEDU (2020).

[28] According to UNESCO (2020), national policies issued by education ministries in South America target home language in 59% of countries, ethnic minorities and indigenous peoples in 56%, gender in 43%, and people with disabilities in 31%; more than half of Latin American and Caribbean countries refer to inclusive education as a priority in their education sector plans or strategies.

3 Data

The empirical analysis in this paper relies on several sources of administrative and survey microdata.

In Section 4 we estimate a difference in differences model using the 2017 Peruvian Census of the Population. The Census includes questions on basic demographics, educational attainment, and employment.[29] We produce aggregates by cohort for each province of birth for the variables of interest. The same variables are also produced separately by gender and by native language.

We use the variable reporting the first language learned as our definition of ethnicity, creating a dummy variable for all individuals that first learned a language other than Spanish.[30] The Census provides 8 suggested options, but, considering responses other than the suggested ones, documents the presence of more than 50 surviving Indigenous languages.[31]

We report summary statistics of the relevant 2017 Census variables in Table 1. Statistics are reported for individuals aged 26 to 75 separately depending on whether the province of birth had a college campus in 2017. We can see that individuals from provinces with one or more college campuses are more likely to learn Spanish as their first language, have higher educational attainment, are more likely to be employed, and to give birth at least once.

We build our treatment variables using information on the time of opening of new campuses. However, administrative records generally report the year of creation of universities, but not the year in which additional campuses admitted their first cohort. First, we use administrative records on the registration of majors to identify the location of university branches.[32] Then, we combine several administrative sources and institutional information from the universities to infer the opening of each additional campus.[33]

In Section 5, we use a novel dataset combining information from high school records,

[29] The complete questionnaire can be found at the following link.

[30] An alternative definition of ethnicity could be constructed using the answer to the question "Given your customs and ancestors, how do you feel or consider yourself?". However, the answer to this question is inherently subjective and potentially affected by the treatment variable. For this reason, we prefer as measure of ethnicity the more objective answer to the question "What is the language that you first learned to speak as a child?".

[31] Appendix Figure 10 shows the percentage of the population that learned an Indigenous language as their first at the province level in 2017.

[32] Similar information can be obtained through the ESCALE system of the Ministry of Education.

[33] A detailed description of the data construction process is included in the Appendix. The distribution of the reconstructed year of entry by type of college ownership is shown in Appendix Figure 11.

Table 1: Summary Statistics of 2017 Census

	N	Mean	SD	Min	Max
Has College					
Age	11729322	44.5	13.0	26	75
Female	11729322	0.52	0.50	0	1
Spanish Native Speaker	11729322	0.86	0.35	0	1
Completed High School	11729322	0.66	0.47	0	1
Attended College	11729322	0.23	0.42	0	1
Completed College	11729322	0.18	0.38	0	1
Employed	11729322	0.60	0.49	0	1
Gave Birth	5978651	0.85	0.36	0	1
No College					
Age	3386324	46.6	13.5	26	75
Female	3386324	0.51	0.50	0	1
Spanish Native Speaker	3386324	0.63	0.48	0	1
Completed High School	3386324	0.44	0.50	0	1
Attended College	3386324	0.10	0.30	0	1
Completed College	3386324	0.085	0.28	0	1
Employed	3386324	0.52	0.50	0	1
Gave Birth	1729881	0.90	0.29	0	1

Notes. Summary statistics for the relevant variables of the 2017 Peruvian Census. The top panel reports the statistics calculated using individuals born in provinces that are treated at some point in time with the opening of a college campus; the bottom panel is calculated using individuals born in "pure control" provinces, that have never had a college campus. Age is restricted between 26 and 75. Further restrictions, leading to the exclusion of some provinces (including the capital Lima), are discussed in Section 4.1. Presence of a college campus is obtained through administrative records, as described in Section 3. *Age* is declared at the time of the Census (2017); *Female* represents gender as reported; *Nat. Spanish Speaker* is a dummy for whether the first language learned as a kid was Spanish; *Completed High School* is a dummy for secondary schooling completion; *Attended College* is a dummy for college enrollment (but not necessarily completion); *Completed College* is a dummy for college completion; *Employed* is a dummy for self reported employment in the week previous to the Census measurement; *Gave Birth* is a dummy for having delivered a living baby (only applies for women 12 or older).

college applications and enrollment. This allows us to track students in the first year of high school through college enrollment. High school records are included in the SIAGIE dataset maintained by the Ministry of Education; similarly, application (successful and unsuccessful) and enrollment data come from the SIRIES dataset. More information about dropout, application, and enrollment rates is reported in Section 5.

We complement this data with three main variables. First, we define the ethnicity of each individual based on the district (approximately correspondent to US census tracts) of birth. An individual is considered from an ethnic minority if born in a district where at least 80% of the population learned a language other than Spanish as their first based on the 2017 Census. Second, we use information on the high school attended by each individual and the dataset on campus locations previously described to identify the closest educational options for each individual. Finally, we collect data on tuition levels: we combine several sources to build a novel dataset with the most detailed tuition information to our knowledge. While cost measures are aggregated at the university level, disregarding differences in costs across campuses, our tuition measure is defined at the campus level by taking the median cost of all the majors offered there. This allows us to have more precise cost information without relying on self-reporting of students in surveys. Details on how the data is collected are available in the Appendix.

Other sources of data used to produce auxiliary evidence, like the National Survey of Households (*ENAHO*), are briefly described as needed along with their analysis.

4 Effects of Campus Openings

4.1 Empirical specification

In order to answer the question of how much the opening of new college campuses increases local enrollment and which groups are most affected by these openings, we use a difference in differences approach. This means comparing changes in enrollment rate in provinces where a college was opened to concurrent changes in other areas. Such focus on *changes* is necessary to address non-randomness in the location of new colleges, which represents the main identification threat to cross-sectional comparisons. This set up suggests a model of the

form:

$$y_{t,p} = \sum_{\tau} \sum_{p'} \beta_{\tau,p'} \mathbb{1}\{p' = p\} \mathbb{1}\{\tau = t - g(p)\} + \psi_t + \mu_p + e_{t,p}$$

where $g(p)$ is the period in which a college opened in province p, and t indexes different cohorts. $\beta_{\tau,p}$ is the parameter of interest: the effect of exposure to the campus opening on the outcome $y_{t,p}$ (e.g. enrollment rates). The term $\beta_{\tau,p}$ highlights the possibility of effects being heterogeneous for different provinces (p) and depending on exposure length ($t - g(p)$).[34] Allowing for heterogeneity and dynamics requires us to avoid standard "two-way fixed effects" (TWFE) regressions, as they have been shown to be problematic in such setups (Goodman-Bacon (2021), Baker et al. (2021)).

The frailties of TWFE come from the inclusion of already-treated groups within the comparison group: if treatment effects are heterogeneous, the TWFE estimator will be biased.[35] Several recent papers have introduced solutions to address these problems, by making sure that only never-treated or not-yet-treated units are included in the comparison group. Similar to the basic difference in differences, these papers rely on parallel trends assumptions to build consistent estimators that do not suffer from the same problems.[36]

We use the estimator proposed in Callaway and Sant'Anna (2020) and provide estimates using only never-treated units, and including not-yet-treated units as well. This estimator is robust to dynamic effects (e.g. increasing with length of exposure to treatment) and effects being heterogeneous across provinces (e.g. because of colleges having different sizes or majors). Standard errors are calculated using the bootstrap procedure suggested in the paper and implemented in the statistical software provided by the authors.[37]

Starting from the full-count individual level data of the 2017 Peruvian Census, we obtain enrollment rates (and other measures) at the age cohort-by-province level. We use the *province of birth* to match individuals to the creation of new campuses. Cohorts will be con-

[34] In our setting, heterogeneous treatment effects are likely to arise from heterogeneity in the capacity of new campuses, or in the population size and area of the province.

[35] Dynamic effects, e.g. when treatment effect is growing as time from the event passes, also lead to misspecification and inconsistency when a constant effect is assumed. This can be addressed by estimating treatment effects relative to event time (Borusyak et al. (2021)). The Appendix provides evidence of the existence of negative weighting in our setting.

[36] See as examples of such new estimators Callaway and Sant'Anna (2020), Borusyak et al. (2021), and Sun and Abraham (2020).

[37] Standard errors are clustered at the treatment level, and critical values in event study graphs are robust to multiple hypothesis testing.

sidered treated if their birth province had a university campus opened by the time they were 18.[38] We only include in the analysis those individuals born in the period 1942 to 1991.[39] The number of openings is 56 for public colleges and 13 for private ones.

We estimate the difference in differences model for everyone and separately for different demographic groups to study the heterogeneity of treatment effects. For each specification, the Appendix reports "event study figures" where we plot treatment effects estimates for each cohort relative to the opening of the campus: these figures allow for a visual inspection of pre-trends and an assessment of the effect's dynamics. In addition, we also divide treatment events in two groups depending on whether the opened campus is public or private. "Event study figures" for these specifications (using the whole population or separate demographic groups) are also reported in the Appendix.

4.2 Results

First, we can see from Figure 4 that we do not find evidence of pre-trends in enrollment and completion. In the same Figure, we also observe that effects are increasing over time, and estimates are statistically significant for individuals 16 and younger, who are still in high school at the time of first admission.

Table 2 reports the average treatment effect for models estimated with the difference in differences estimator described in Section 4.1. The outcomes are whether an individual ever enrolled in a college, and whether she completed it. In this Table, we separately report average treatment on the treated in the short and long run. Short-run effects are those for cohorts in school (aged 6-16) at the time of first treatment, while long-run ones include all treated cohorts.[40]; in the rest of the paper we will focus on the effects of public openings for school-aged cohorts only, and provide the effects on all cohorts and private openings in the Appendix.

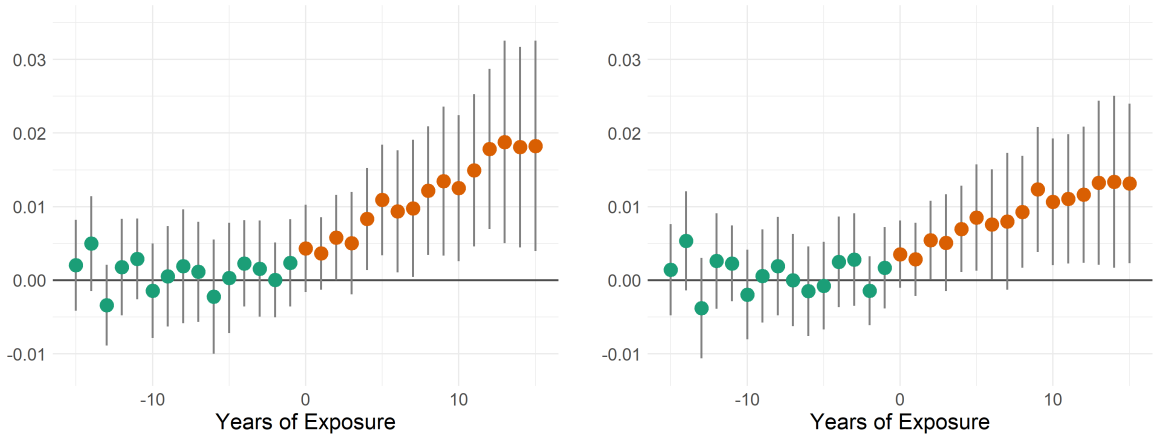
The first row of Table 2 shows that college openings significantly increased enrollment

[38] Section 3 and the Appendix describe the process used to construct the dataset with the opening of campuses.

[39] Limiting our analysis to cohorts born after 1942 helps to reduce bias from differential mortality rates (in 2017 an individual born in 1942 would be 75 years old). Limiting our analysis to cohorts born before 1992 allows time for the outcomes of the youngest individuals to fully realize (in 2017 an individual born in 1991 would be 26 years old).

[40] Contrary to TWFE, the estimator in Callaway and Sant'Anna (2020) separates estimation from aggregation of effects, allowing us to estimate different average effects.

Figure 4: Effects of College Openings for College Enrollment (*left*) and Completion (*right*)



Notes. Difference in differences estimates obtained using the estimator from Callaway and Sant'Anna (2020), as described in Section 4.1. Event study graphs for the effect of the opening of any kind of college campus on enrollment (left panel) and completion (right panel). Source: 2017 Peruvian Census of the Population.

Table 2: Diff-in-Diff: ATT Estimates by College Type

	College Enrollment		College Completion	
	Public	Private	Public	Private
Short Run				
	.0104***	.0156**	.008***	.0133*
	(.0023)	(.0073)	(.002)	(.0074)
Long Run				
	.0224***	.0427***	.016**	.0309***
	(.0084)	(.0073)	(.0072)	(.0058)
Baseline	.105	.091	.1274	.108
# Treated Provinces	56	13	56	13
# Control Provinces	113	113	113	113

Notes. This table shows the estimated ATT for the difference in differences model using the estimator proposed in Callaway and Sant'Anna (2020), as described in Section 4. The dependent variables *College Enrollment*, and *College Completion* are dummy variables. Different columns report estimates for two different treatment definitions for each dependent variable: in the first model (indicated by "Public") we define as treated a cohort aged 18 or younger at the time of first opening of a *public* college campus; in the second (indicated by "Private") only openings of *private* college campuses are considered. Standard errors (in parentheses) clustered at the treatment level are calculated using the bootstrap procedure developed in Callaway and Sant'Anna (2020) and implemented in the provided software. Baseline values are dependent variables' means for 20-year-old individuals at the time of opening of a new campus. Stars represent statistical significance of the single hypothesis test. *** p<0.01, ** p<0.05, * p<0.1.

and completion at universities, with a smaller effect for public institutions than for private ones. Larger effects for private openings might be due to a preference for private institutions over public ones, or, most likely, to differences in capacity constraints and selectivity. Private universities can expand their capacity faster than public ones that are financed only through public funds[41] and are unable to expand in response to an increase in applications.

Table 3 estimates the same model on subgroups of the population, according to their ethnicity and gender. We can see that individuals in the ethnic minority experience very small increases following the opening of a public or private campus. This implies a widening of the ethnicity gap. This is shown both by using the gap as outcome and by the rejection of the hypothesis of the effects on minority and majority being the same. The heterogeneity by gender is not as strong, but is consistent with women being more responsive on the enrollment margin, as found previously in related situations (e.g. see Carrell and Sacerdote (2017)).

The effects of new college campuses likely come through the reduced cost of attendance thanks to proximity and changes in the probability of admission for those who apply. It is reasonable to assume that some groups benefit more than others from closeness to educational options, and that some students will be better able to take advantage of the system to be admitted. In Section 5, we will focus on disentangling how this two channels explain the heterogeneous effects we observed. We will use recent data following students through their high school, application decisions, and enrollment outcomes. This data includes information about important demographics that are unavailable in the census, and also reports instances where students unsuccessfully applied to college. Tooled with estimates about the importance of each channel, we will be able to simulate the impact on enrollment gaps of different policies.

Robustness

As a first robustness exercise for our results, we replicate our analysis using a different definition of treatment, based on distance from college rather than on the province of birth. We consider treated those individuals who had a college available within 40km (25mi) from

[41] The amount of funds for the year is agreed between each public university and the Ministry of Economics and Finance. Historically, these funds have been adjusted based on previous years' funding and enrollment numbers; only recently performance indicators have been introduced in the process.

Table 3: Diff-in-Diff: Heterogeneous Short Run Effects of Public Entry

	Ethnicity			Gender		
	Majority	Minority	Gap	Female	Male	Gap
Enrollment						
	.013***	.0061	.0097*	.0116***	.0092***	.0024
	(.0031)	(.0038)	(.0051)	(.0028)	(.0029)	(.0037)
p-value $\text{ATT}_1 = ATT_2$	0.074					
Completion						
	.0111***	.0041	.0099*	.0091***	.0069**	.0021
	(.0027)	(.0035)	(.0051)	(.0023)	(.0029)	(.0031)
Baseline Enrollment	0.138	0.040		0.088	0.129	
Baseline Completion	0.119	0.035		0.075	0.111	
# Treated Provinces	56	28	28	56	56	56
# Control Provinces	112	65	64	110	100	110

Notes. This table shows the estimated ATT for the difference in differences model using the estimator proposed in Callaway and Sant'Anna (2020), as described in Section 4. The dependent variables *College Enrollment*, and *College Completion* are dummy variables. The ATT is obtained by averaging the dynamic effects estimated on cohorts aged 6-16 at the time of opening. Estimation is carried out using only Not-Yet-Treated provinces as comparison group. Different columns report the estimated ATT for different subgroups of the population. All columns report estimates where we define as treated a cohort aged 18 or younger at the time of first opening of a public college campus. Standard errors (in parentheses) clustered at the treatment level are calculated using the bootstrap procedure developed in Callaway and Sant'Anna (2020) and implemented in the provided software. We exclude provinces with less than 2500 individuals for the subgroup; specifications where *Gap* is the outcome require 2500 individuals for each subgroup. Gaps are calculated by subtracting the share of the minority (males) from that of the majority (females). Baseline values are dependent variables' means for 20-year-old individuals at the time of opening of a new campus. The p-value for the difference of estimates is obtained through a bootstrapping procedure. Stars represent statistical significance of the single hypothesis test. *** p<0.01, ** p<0.05, * p<0.1.

their district of birth by age 18.[42] To address the possibility of spatial spillovers, we introduce buffer areas and exclude from the control group individuals born between 40 and 60km from the college location. As can be seen in Appendix Tables 7 and 8, all results are comparable to the ones previously described.

Another way to verify the possibility of spillovers that might downward bias our estimates is to estimate the treatment effect for different distance groups. In Appendix Figure 13, we see that enrollment effects are largest among individuals born within 20km from the college location, and no effect is found for those born farther than 40km. Given the average size of Peruvian provinces, it is unlikely that spillover effects will bias our original estimates.

One potential concern is that parallel trends only hold conditional on relevant covariates: for example, it might be that trends depend on the location of the province or on its size. Callaway and Sant'Anna (2020) provide an alternative estimator with the double robustness property found in Sant'Anna and Zhao (2020): estimates are unbiased as long as either a propensity score or an outcome regression models is correctly specified. Event study plots for this specification show no evident pre-trends with effects similar to the ones of our main specification when using controls for latitude, longitude, and population.

Other outcomes

The census also allows us to observe a number of other outcomes that are potentially affected by college openings and enrollment. Table 4 reports estimates of the average treatment effects on community college enrollment, high school completion, fertility, employment and migration.

We find that enrollment at community colleges does not appear to be affected in a statistically significant way. Given that estimated effects are ranging between 0.09 and 0.21$p.p.$, we consider the impact on community college enrollment to be negligible.[43] Notice that high school completion appears to be positively affected, even though the estimates are not significant at standard levels: this is consistent with the predictions in Eisenhauer et al. (2015) and in contrast to the theory and findings in Bedard (2001). In the same way, fertility and mi-

[42] Peruvian district can be considered comparable to US zip code areas.

[43] The confidence interval for all openings has a lower bound of $-0.38p.p.$, or about one fifth of the point estimate for college enrollment.

Table 4: Other Outcomes: Diff-in-Diff ATT Estimates by College Type

	Public	Private
Community College	.0009	.0021
	(.0027)	(.0054)
Completed HS	.0069	.0173
	(.0056)	(.0185)
Employed	.0096*	.0027
	(.0049)	(.0063)
Had Kids	-.0045	-.0028
	(.0029)	(.0087)
Migrated	-.0058	.0112
	(.0057)	(.0116)

Notes. This table shows the estimated ATT for the difference in differences model using the estimator proposed in Callaway and Sant'Anna (2020), as described in Section 4. The dependent variables *HS Completion*, *Has Kids*, *Employment*, and *Migration* are dummy variables. Estimation is carried out using only Not-Yet-Treated provinces as comparison group. Different rows report the estimated ATT for different subgroups of the population. Different columns report estimates for two different treatment definitions for each dependent variable: in the first model (indicated by "Public") we define as treated a cohort aged 18 or younger at the time of first opening of a *public* college campus; in the second (indicated by "Private") only openings of *private* college campuses are considered.. Standard errors (in parentheses) clustered at the treatment level are calculated using the bootstrap procedure developed in Callaway and Sant'Anna (2020) and implemented in the provided software. Stars represent statistical significance of the single hypothesis test. *** p<0.01, ** p<0.05, * p<0.1.

gration appear to be somewhat reduced. Finally, we find positive and statistically significant (at 10% threshold) effects on employment, only for public openings. This is not surprising as Peru has recently undergone a licensing process that led to the closure of one third of universities due to low quality standards, almost all of them private institutions.

5 Model of Education Demand

The analysis in Section 4 shows that (i) enrollment and completion increased as a consequence of the opening of new universities in the proximity of students, and (ii) the increase was concentrated among individuals belonging to the ethnic majority and women. However, the role of different channels in shaping these results is unclear: do minority students benefit less from colleges opening because they don't care about distance to college? Or, instead, they become more likely to apply but have lower chances of admission? The answer to these questions is necessary to devise better policies. If some groups of students are systematically less likely to apply, policymakers might want to provide financial aid to make going to college a more appealing (or feasible) option. On the other hand, if entrance exams represent the relevant bottleneck, affirmative action policies would represent a more effective tool.

The patterns observed in the previous Section call for a model where different individuals value and react differently to changes in the distance from educational options. Similarly, having identified the admission process as a potentially important determinant of who eventually enrolls, we need to reproduce its effects and allow policy to affect it. These considerations lead us to build a discrete choice model of demand for education and college admissions, tailored to the Peruvian context.

5.1 Model setup

We combine several administrative records to obtain a dataset containing detailed individual information on demographics, ability proxies, college applications and enrollment.[44] This

[44]Census data only reports few variables to characterize individuals: the dataset that we use in this section provides us with information on parental educational attainment, place of birth, high school attended and GPA, among other variables that have been shown to be important for educational attainment in other studies. Additionally, not observing application behavior and admission results at the individual level would require imposing

will allow us to estimate separately admission probabilities and preference parameters for a cohort of Peruvian students expected to graduate high school in 2018. These students are tracked during high school and in their application and enrollment decisions over the period 2013-2020, as described in Section 3.

Individuals in our model can reach four "levels" of educational attainment: public university degree (giving utility u_{public}); private university degree ($u_{private}$); high school diploma (u_{hs}); or high school dropout ($u_{dropout}$). To reflect the fact that a student whose desired outcome is to attend a (selective) public university ($u_{public} \geq u_j, \forall j \in \{private, hs\}$) might not be admitted and have to choose another option, we modify a standard conditional logit by introducing uncertainty over admissions. Given information about both applications (successful and not) and enrollment, we know the favorite option of all individuals[45] and the "second-choice" option for those who were not admitted to a public university. This provides more information than in a setting where only the realized outcome or only the application decision is known. We further specialize our model by making the attainment choice dynamic, e.g. a student whose application to a public university has been rejected cannot choose to drop out of high school.[46]

Figure 5 shows a graphical depiction of the model. Individuals make two sequential choices: first, they choose whether to drop out of high school based on the expected utility from graduating given by the maximum utility of the subsequent choices and on individual preference shocks – utility of post-graduation options is not observed at this time.[47] Then, those students who did not drop out observe their preference shocks for all remaining options and decide whether to apply for a public college, enroll in a private one, or not pursue any tertiary education option. If enrollment at a public college represents the preferred option, an application is made that results into admission with probability γ_i. If the applicant is not admitted, she has to choose between the other two remaining options. Note that a stu-

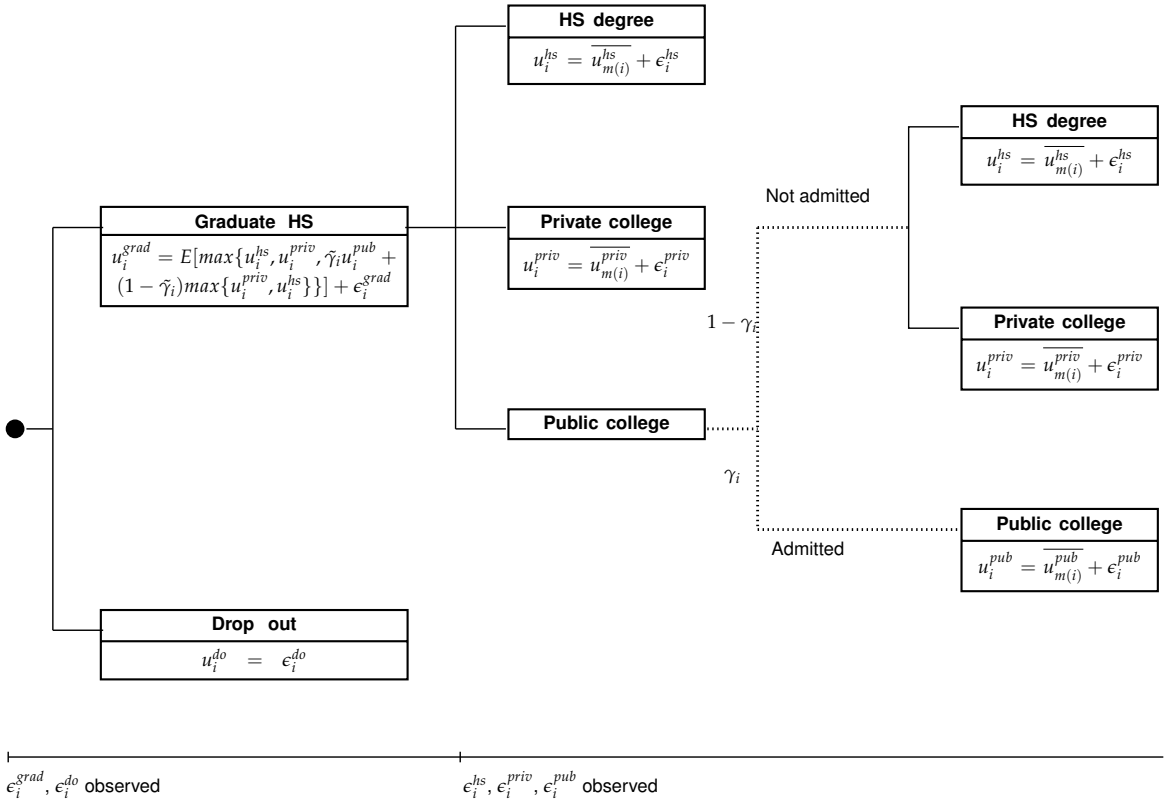
additional, strong assumption on the data.

[45]Here, we assume that a student who applies to a public university would receive the most utility from attending it. If applications are costless, all individuals whose favorite choice is enrollment in a public university would apply, even with very small probabilities of admission.

[46]One of our counterfactual policies involves an improvement in secondary schooling quality, which is likely to affect dropout decisions, potentially leading to important welfare consequences. Additionally, high school dropout decisions might be affected differentially by college openings for some subgroups of the population.

[47]Calculating expected utility requires that students use their beliefs over the probability of admission to public colleges, which we indicate as $\tilde{\gamma}_i$.

Figure 5: Model for education demand



Notes. Graphical representation of the estimated model. Students first choose whether to drop out or not, without observing shocks to their individual utility from other options. Upon graduation, all shocks are observed and students choose whether to apply for a public college, enroll in a private one, or not pursue any tertiary education option. Students who apply to public college are admitted with probability γ_i. If not admitted, they will chose between the two remaining options: enrolling in a private one, or not pursuing any tertiary education option.

dent who applies to public college will always choose to enroll when admitted, as the other options were available to her in the first place.

Finally, we make assumptions about the admission process and indirect utility. As previously highlighted, and given the findings of Section 4 it will be crucial to allow for meaningful heterogeneity along the studied dimensions. Before describing the assumed specifications, it is useful to define some common variables and labels. We group GPA in three bins: low (lowest quartile), medium (between 25th and 50th percentile), and high (above median). Socioeconomic status (SES) is considered high when at least one of i's parents completed high school. Ethnic minority status is assigned for individuals born in districts where at least 80% of individuals learn an indigenous language as kids, according to the 2017 Cen-

sus. High school quality is represented by a binary variable for having above median value added. We use a national standardized test to estimate value added for each high school in the country: details about the data and value added estimation are reported in the Appendix. Proximity is a binary variable taking value of 1 when option $j \in \{public, private\}$ is available in the same province where i attended high school. $d(i)$ is the region (*departamento*) where i attended high school; $p(i)$ is the province where i attended high school; $m(i)$ defines the set of individuals with the same observable characteristics (SES, GPA, gender, ethnicity, location, high school quality) as i.

Probability of admission to public college for individual i is

$$\begin{aligned}\gamma_{m(i)} = \lambda_1 GPA\_high + \lambda_2 GPA\_low + \lambda_3 high\_SES + \lambda_4 minority + \lambda_5 female \\ + \rho_1^i hs\_quality + \rho_2^i proximity + \lambda_6 \mathbb{1}\{p(i) = Lima\} + \psi_{d(i)} + e_{m(i)}\end{aligned} \tag{1}$$

$$\text{with} \quad \rho_k^i = \rho_k + \textstyle\sum_h \rho_k^h D^{ih}$$

and where D are dummies and $h \in \{high\_SES, minority, female\}$. The ρ^i terms allow for heterogeneous impact of high school quality and proximity on the probability of admission.

We assume that utility of individual i from choice $j \in \{hs, public, private\}$ equals

$$\begin{aligned}u_{ij} = &\beta_1^i proximity_{ij} + \beta_2^i hs\_quality_i \\ &+ \eta_1^j high\_SES_i + \eta_2^j hs\_quality_i + \eta_3^j female_i + \eta_4^j \\ &+ \theta_1 tuition_{ij} + \theta_2 wage\_premium_{ij} + \theta_3 GPA\_high_i + \theta_4 GPA\_low_i + \phi_{p(i)} + e_{ij}\end{aligned} \tag{2}$$

$$\text{with} \quad \beta_k^i = \beta_k + \textstyle\sum_h \beta_k^h D^{ih} \quad , \quad \eta_k^j = \eta_k + \eta_k^{hs} \mathbb{1}\{j = hs\} \quad \text{and} \quad e_{ij} = \xi_{m(i),j} + \epsilon_{ij}$$

where D^{ih} are dummies for ethnicity, gender, and SES. $\phi_{p(i)}$ are fixed effects for the province where i attended high school. We normalize utility from the outside option, $u_{i,dropout} = \epsilon_{i,dropout}$. All ϵ_{ij} errors are assumed to be i.i.d. EV1 (logit) errors. $\xi_{m(i),j}$ can be interpreted as unobservable characteristics at the level of the group defined by $m(i)$. Wage premiums in each province are calculated relative to dropout wages using the household survey ENAHO.

It is worth noting some important simplifying assumptions. Private universities are assumed to be non-selective: this is correct for the vast majority of private colleges, with few ex-

ceptions that are primarily located in Lima. Indeed, according to the administrative dataset SIRIES, 78% of applications to private colleges in 2017 were successful, against 20% for public colleges. We also assume that applications to universities are free, but their outcome is final: applications fees for the median public college were 228 Soles in 2017 or USD68 (1% of GDP per capita), according to the 2017 ENUE survey, while the corresponding amounts for private colleges were 120 Soles (USD37).[48] While applying multiple times is possible, most students only apply once[49], and we calculate the probability of admission considering the possibility of multiple attempts. This simplifies the model by reducing the choice space[50], while assigning appropriate probabilities of admission to students. Finally, notice that individuals do not know their preferences regarding tertiary education until after high school graduation, and the expected utility of graduating from high school is correct only on average. These assumptions appear reasonable given the extensive literature documenting the potential of informational interventions in education and the dispersion of beliefs over returns from college (see, e.g., Hastings et al. (2016)).

5.2 Estimation

Before estimating preferences, we predict the individual probability of admission γ_i. Because we observe the outcome for each individual who applied, we can use it as dependent variable in a linear model including individual characteristics. In this case, we can use OLS to estimate model (1). Estimates from this specification and some alternative ones are reported in the Appendix.[51] The predicted probability of admission will be called $\hat{\gamma}_i$.

To estimate utility parameters, we use a two-steps approach, similar to those used by Hastings et al. (2017) and Abdulkadiroğlu et al. (2020). The first step will recover mean utilities, and the second will use them to estimate the parameters of interest in the indirect utility function. Intuitively, the first step takes care of individual level unobservables contained in

[48] This information is calculated using the Ministry of Education's ENEU survey of college students and is not necessarily representative of the median cost of all applicants.

[49] According to the census of college students of 2010 (CENAUN), 43% of students applied to only one university and 28% to two; 73% of college students applied just once to the college they attend (15% applied twice).

[50] Modeling the possibility of applying after being rejected would add another choice to the "Not admitted" node.

[51] One potential source of bias is selection on unobservables of students that apply to college – e.g. students with low test-taking abilities might not apply at all. While possible, this is unlikely to importantly affect the estimates given the inclusion of several ability measures.

ϵ_{ij}, so that only $\xi_{m(i),j}$ is relevant for the second step.

In the first step, we use Maximum Likelihood Estimation to obtain mean utilities[52] for each option. Mean utilities are estimated for each group of individuals that has the same observable characteristics, indexed by $m(i)$: out of 9408 potential groups, 2365 are populated by individuals and have non-zero shares for all options.[53] At this step, we can include information about first and second choice for individuals who apply to public college and are rejected. In the Appendix, we describe in detail the Likelihood function used for estimation. Two key elements should be highlighted here: the first is that we will plug into the Likelihood function the $\hat{\gamma}_i$ we obtained through our linear probability model; the second is that, in addition to mean utilities, we are able to recover the beliefs about the probability of being admitted to college, $\tilde{\gamma}_{m(i)}$.[54] As is usual for conditional logit models, we normalize the utility of one choice, dropping out of high school, to have zero mean – coefficients will be identified relative to that option. Relatedly, we also set the variance of all errors (scale parameter) to be equal to 1.

In the second step, we project regressors and instruments on the estimated mean utilities to recover the parameters of interest of model (2). We use an IV approach to obtain estimates for tuition's parameter, θ_1: because tuition is set endogenously by private universities, we need to use an exogenous instrument in order to recover a consistent estimate. We use the interaction between distance from the closest public college, the difference in wage between college and high school graduates, and an indicator for private colleges as an instrument. Intuitively, the higher the college premium, the more private colleges might be able to charge; distance from the closest public college proxies competition. When returns are high and competition low, private colleges will charge higher tuition. The exogeneity assumption needs to be holding conditional on province fixed effects and local returns to college, among other controls. Then, the second stage model equation will be

$$\overline{u^j_{m(i)}} = \theta_1 \widehat{tuition}_{ij} + \boldsymbol{X_{ij}\beta} + \xi_{m(i),j}$$

[52] Given the notation used in model (2), mean utility is equal to $\bar{u}_{m(i),j} = u_{ij} - \epsilon_{ij}$.

[53] We only need non-zero shares for dropout, private college enrollment, no enrollment, and public college application (not necessarily admission).

[54] $\tilde{\gamma}_{m(i)}$ is identified by variation in the probability of admission to public colleges and their utility.

, where $\widehat{tuition}$ is the tuition predicted in the first stage regression.

5.3 Estimation Results

Figure 6 shows how the distribution of the predicted probability of admission for each demographic cell, $\hat{\gamma}_i$ aligns with the observed ones, as reflected by the R^2 of 0.73 for the chosen model of column 4 in Table 9 in the Appendix. We can see that some groups appear to have probabilities equal to 0 or to 1 in the data, due to their small size; this issue is solved by the linear probability model. We impose the constraint that predicted probabilities are strictly between 0.01 and 1, which is binding for 1.1% of the individuals in the sample with negative probabilities (the minimum predicted probability is -5.6%). In Appendix Table 10, we can see that being close to a public college increases the probability of admission by almost $5p.p.$. Further discussion about different specifications for the LPM are included in the Appendix.

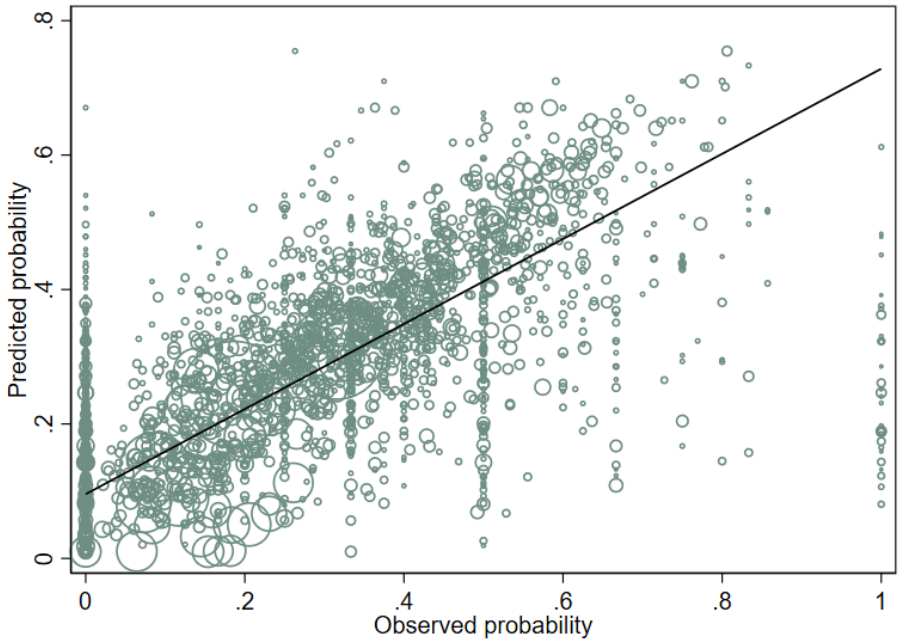
In Table 5 we report the results from the second step of the estimation. Columns 1 to 3 report results using different definitions of distance/proximity. Tuition is negatively related to utility as predicted by standard demand theory, suggesting that the instrument is effective in removing the endogenous relation between prices and quality or demand shocks.[55] We find that students value proximity at about \$500 or 1/4 of average private tuition. When estimating the effect separately for high- and low-SES students, we can see that the latter value proximity about twice as much as the former. In column 3, we see that changing the definition of distance from presence in the same province of the high school attended to kilometer distance from the closest option does not affect this conclusion. The instrument appears to be relevant, as shown by the Kleibergen-Paap F-Stats around 70 and well above the standard rule of thumb. Standard errors are clustered at the province level, as in Hastings et al. (2017).

In order to assess the fit and validity of the estimated model, we compare the shares of each choice[56] in the data and the predicted shares from the model. The results for all individuals and for different subsamples are reported in Table 6. We can observe that the model does a good job at predicting choice shares overall, and is able to sensibly reflect

[55] Tuition is measured in thousands of Soles, and ranges between 7 and 9 for private colleges. One thousand Soles corresponds to approximately \$250.

[56] We report dropout, enrollment at public or private colleges, or high school completion as outcomes. Not everyone who applies is allowed to subsequently enroll in a public university, as discussed previously.

Figure 6: Probability of Admission, Observed and Predicted



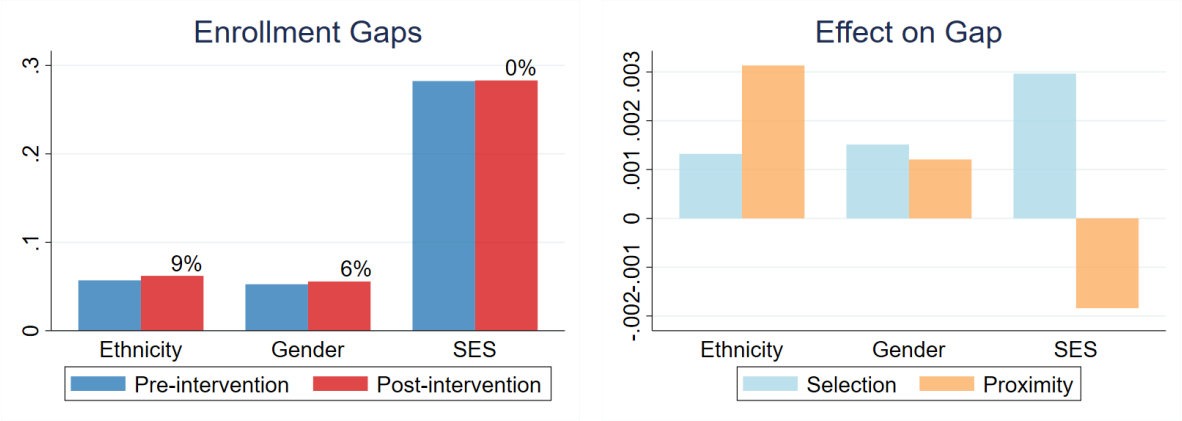
Notes. The x-axis of this figure reports the fraction of students that was admitted to a public college among those who applied, for each demographic group. The y-axis represents the predicted probability of admission predicted by Model (1). The size of each marker represents the number of individuals in the demographic group. The reported line represents the linear fit of predicted on observed probability of admission. Information on applications and admissions is contained in the SIRIES dataset.

Table 5: Estimated Utility Parameters

	(1)	(2)	(3)
Tuition	-0.195***	-0.193***	-0.172***
	(0.0345)	(0.0351)	(0.0379)
Proximity	0.383***		
	(0.108)		
Proximity (Low SES)		0.440***	
		(0.111)	
Proximity (High SES)		0.249	
		(0.157)	
Distance (Low SES)			-0.00881***
			(0.00161)
Distance (High SES)			-0.00442**
			(0.00220)
Observations	1377858	1377858	1377858
Kleibergen-Paap F-stat	67.94	69.00	70.00

Notes. This table shows the estimated coefficients for model (2) using an instrumental variable approach, as described in Section 5. The dependent variable is the mean utility for each available option (graduating high school with no further education, attending a public college, attending a private college) estimated in the first step through Maximum Likelihood. All models also include dummies for demographic characteristics (gender, parental education, minority status, above median GPA, lowest quartile GPA, high school quality, province of attended high school) and several interactions. The instrument used for the estimation is the interaction of a dummy for the private college option, a measure of local college premium, and distance from the closest public college campus. The number of different demographic combinations for which utility estimates (our dependent variable) are available is 2365; 191 different provinces are represented in those groups. Standard errors (in parentheses) are clustered at the province level. *** p<0.01, ** p<0.05, * p<0.1.

Figure 7: Counterfactual Gaps and Channels



(a) Counterfactual 1: Gaps

(b) Counterfactual 1: Channels

Notes. Gaps in enrollment before and after the policy implementation by ethnicity, gender, and SES (left); effects on gaps of selectivity and proximity channels. The sample is composed of provinces where no college is currently available. Policy implemented: a new public college is opened in each province. Gaps after the implementation of the policy are obtained by letting the probability of admission γ_i and utility u_{ij} change to reflect increased proximity to a public college. Channels are separated by only letting one variable at a time change.

heterogeneity of different groups.

5.4 Counterfactual exercises

The estimated model allows us to (i) simulate the effects of alternative policies on enrollment, and (2) quantify the role of admission criteria and heterogeneous preferences in determining the observed outcomes. When simulating our policies, we will let both the probability of admission (γ_i) and utility (u_{ij}) adjust in response. To separate these two channels we will compare the changes in gaps that would happen if only one of the two key elements at a time is allowed to respond to policy. In this Section, we will study two counterfactual policies: opening new college campuses, and a reduction in upstream inequalities in the access to high-value added (VA) schools.

First counterfactual: opening public colleges in currently underserved areas

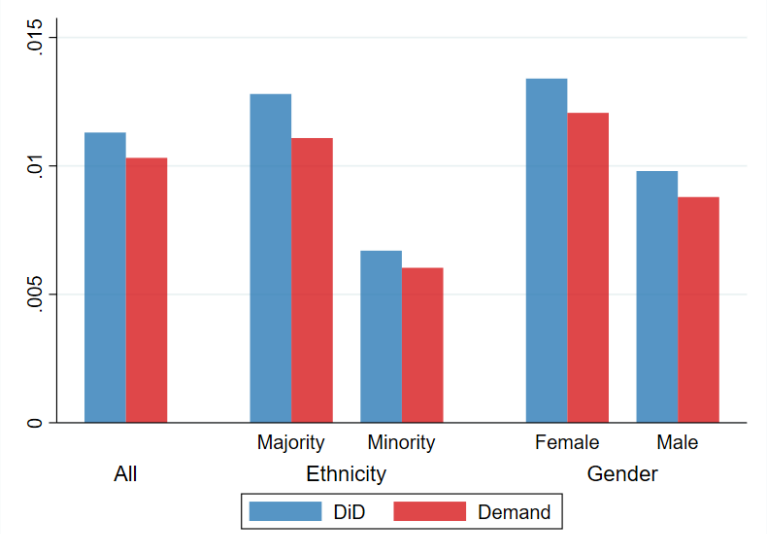
In the first counterfactual, we simulate the opening of public college campuses in provinces that do not have any. This exercise is similar in spirit to the natural experiment in Section

Table 6: Model Fit of Outcome Shares

		Data	Model
All			
	Dropout	0.173	0.193
	Public College (Admitted)	0.086	0.082
	Private College	0.253	0.182
	High School	0.489	0.542
	N	459642	459642
Female			
	Dropout	0.144	0.158
	Public College (Admitted)	0.089	0.089
	Private College	0.292	0.216
	High School	0.476	0.538
	N	219976	219976
Minority			
	Dropout	0.132	0.14
	Public College (Admitted)	0.076	0.039
	Private College	0.092	0.107
	High School	0.7	0.714
	N	31715	31715
High SES			
	Dropout	0.128	0.15
	Public College (Admitted)	0.112	0.121
	Private College	0.386	0.277
	High School	0.375	0.451
	N	225174	225174

Notes. The first column of this table reports the shares of individuals choosing each available option: dropping out of high school, graduating from high school with no further education, and attending a private or public college. We then compare these shares to those predicted by the estimated model in the second column. We report shares from the data and the model for the whole population studied and for three subgroups – women, students with low GPA, and with parents who completed high school.

Figure 8: Effects of Campus Openings – DiD vs Model Predictions



Notes. Difference in differences estimates compared to discrete choice model prediction for the effect of public college openings. The prediction sample is composed of provinces where no college is currently available. Policy implemented: a new public college is opened in each province. Estimates reported for the difference in differences model are for school aged individuals at the time of opening.

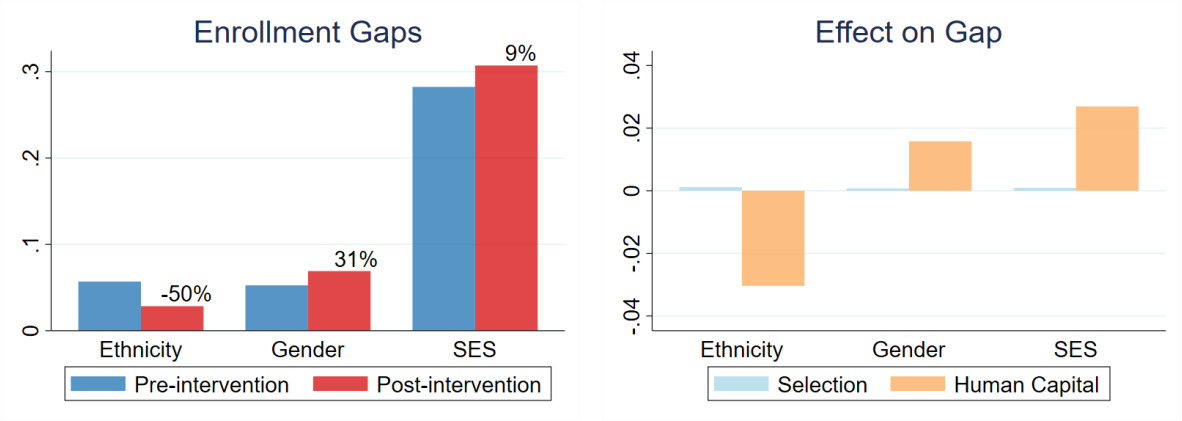
4.[57] The left panel in Figure 7 shows the gaps before and after the opening of colleges. We can see that gaps increase for ethnicity and gender by 9% and 6%, respectively. This is aligned with the difference in differences results. However, the socioeconomic status gap, unobserved in Census data, does not change: that's despite the results in Table 5 showed higher benefits from proximity for low-SES students. In the right panel of Figure 7, we see that while the proximity channel works to decrease the SES gap, its effects are offset by the selectivity channel. In fact, the selectivity channel increases gaps along each of the studied dimensions.[58]

This counterfactual exercise also provides an opportunity for validation of our model. After simulating the opening of new college campuses, we can compare the predicted changes

[57] Two crucial differences are the time frames (1960-2009 for the difference in differences, and 2017 for the current exercise), and the geographical areas, because provinces currently without any college are necessarily not treated in the quasi-experiment.

[58] As an illustrative example of how the selectivity channel can offset the proximity effect, suppose that 50% of high-SES students apply to college with admission probability of 40%, while 30% apply from low-SES with admission probability 30%. This would lead to an $11 p.p.$ enrollment gap between the two groups. If opening a college increases the probability of admission uniformly by $10 p.p.$ and increases application rates by $10 p.p.$ for high-SES students and $15 p.p.$ for low-SES ones, the enrollment gap would increase to $12 p.p.$.

Figure 9: Counterfactual Gaps and Channels



(a) Counterfactual 2: Gaps

(b) Counterfactual 2: Channels

Notes. Gaps in enrollment before and after the policy implementation by ethnicity, gender, and SES (left); effects on gaps of selectivity and proximity channels. The sample is composed of provinces where no college is currently available. Policy implemented: below-median-VA high schools are improved to the level of above-median-VA high schools. Gaps after the implementation of the policy are obtained by letting the probability of admission γ_i and utility u_{ij} change to reflect increased proximity to a public college. Channels are separated by only letting one variable at a time change.

in enrollment for different groups to the estimated effects from Section 4.[59] Figure 8 shows the results of this comparison.[60] We can see that the predictions for the overall population are very similar to the estimates of Section 4, and the model produces remarkably similar heterogeneity along the gender and ethnic dimensions: as our main focus regards the effects on different groups in the population, being able to capture their different reactions to policy is particularly important. This exercise gives us confidence in generalizing our conclusions about mechanisms to the previous part of the analysis.

Second counterfactual: reducing inequality in High School Value Added

The second counterfactual policy addresses a central point in the meritocracy-equity tension: if all students had access to the same resources before applying to college, we would expect more equal outcomes. We focus on access to high-quality secondary education. In our data,

[59] It is worth remarking that, differently from other related papers, we are not using the estimates of Section 4 in the estimation of our discrete choice model. Doing so would make the current exercise substantially less meaningful.

[60] To reflect the characteristics of the data used in the current section, we report the estimated effects of the opening of a public college on school aged students.

we observe that 85% of minority students attend below-median VA high schools; the same proportion for low-SES students is of about two thirds.[61] We simulate an improvement in the quality of instruction that provides everyone with the same quality of current above-median high schools. While this represents a massive investment on secondary education, the primary goal of our exercise is to highlight the importance of prior advantage in shaping inequality in enrollment.[62,63]

In the left panel of Figure 9, we report the predicted gaps before and after the reduction in high school quality dispersion. This policy decreases by 50% the ethnicity gap, reflecting the highlighted disparity they face in accessing high-VA schools. However, the policy also increases the gender gap by 31% and the SES gap by 9%. These last results are explained by two factors: first, the inequality in access to better schools is less strong for men and low-SES than for minority students; second, women and high-SES students benefit more from a high-VA high school. In Appendix Table 5, we can see that high school quality interacted with dummies for women or high SES produces positive and significant coefficients. The right panel of Figure 9, once again shows that selectivity increases gaps, even though its effects are now much smaller relative to those induced by the increase in human capital from improving secondary schooling. The human capital appear to only reduce the ethic gap, while it has positive effects along the gender and SES dimensions.

Our simulations have some limitations that are relevant for the interpretation of results. First, we do not model the pricing decision of private colleges: this means that private colleges are not allowed to change tuition in response to the implemented policies and changes in demand. The focus on provinces without private college campuses makes it less likely that private colleges in other provinces would adjust tuition sizably. The direction of the bias coming from holding tuition fixed depends crucially on the sensitivity of various groups to monetary costs.

[61] To keep predictions comparable, we use the same sample as in the first counterfactual exercise. This means focusing on provinces that do not currently have any college available.

[62] Heckman and Mosso (2014) review evidence about the importance of early life conditions in shaping relevant outcomes. While it is likely that access to relevant resources diverges before high school, our data does not allow explorations of inequality at earlier ages.

[63] Equalization could also be achieved through a reallocation of teachers. Bobba et al. (2021) shows that a budget-neutral redesign of incentives for teachers' mobility could ensure that half of rural schools are staffed with a teacher who is at least as competent as the average teacher in urban areas, while they estimate that complete equalization could be achieved with less than $5 millions.

Second, we do not model credit constraints. Table 11 in the Appendix compares how demographics relate to beliefs and real probability of admission: if beliefs were exactly correct and the model perfectly specified, we would expect them to relate in the same way. Credit constraints are one factor that might explain part of the difference found. As most coefficients have the same sign and comparable magnitudes, we think that credit constraints should not impact our conclusions. Third, the model of γ_i is invariant to changes in demand: this means that, if we have an overall increase in demand (as happens for our second counterfactual), the total number of admitted students might be higher than current capacity constraints. This means that our second counterfactual should be interpreted as a reduction in upstream inequality joint with an expansion in capacity to accommodate the increased demand.

Finally, we assume that students' behavior does not react to the changing environment along unmodeled dimensions, such as the use of private tutoring (e.g. see Chatterjee et al. (2020) for primary education). This possibility cannot be easily ruled out.

6 Conclusions

How scarce college seats are allocated has been a debated topic for a while. Similar to how the University of California system grew from 2 to 9 campuses offering college curricula between 1943 and 1965 to accommodate demand (Stadtman (1970))[64], developing countries have been expanding the supply of higher education at a geographical level and increasing the capacity of existing campuses in the last few decades (e.g., see Ferreyra et al. (2017)). Increasing supply, however, has not made the allocation problem any less salient.

In fact, gaps in access to post-secondary education are still a recognized problem in developed countries like the US (see, for example, Bailey and Dynarski (2011), Hanushek et al. (2020), and Chetty et al. (2020)), and governments of developing countries like Peru have raised concerns of stratification and made equitable access a primary goal (Ministerio de Educación (2020)).

In this paper, we have observed the consequences of meritocratic admission criteria for the allocation of college seats in a context where initial resources, such as high-quality sec-

[64]Recent news have highlighted plans to further expand the system in the next 10 years, see https://www.latimes.com/california/story/2021-10-01/uc-seeks-enrollment-hike-to-meet-college-admission-demand .

ondary education, are unequally distributed. We have shown that in the case of Peru, an expansion of the higher education system through the creation of new campuses has led to greater increases in college enrollment and completion for more advantaged students, thereby increasing attainment gaps. Building and estimating a discrete choice model of demand for education, we have studied the channels determining this outcome. We show that while proximity is valued by low-socioeconomic status students the most, the selection process offsets its equity benefits and leads to increases in pre-existing gaps along all measured dimensions. Advantaged groups are better able to take advantage of proximity to college and increase their chances of admission after college campuses are opened due to complementarities between proximity of educational institutions and demographics of the advantaged groups.

It is important to note that meritocratic selection criteria do not necessarily produce unequal outcomes, just like non-meritocratic ones do not necessarily worsen allocation. In a world where initial resources (like family wealth, networks, educational opportunities) are equally distributed and talent is not prerogative of one group, meritocratic rules might lead to the best allocation *and* equal access to education; because talent is more equally distributed than access in poorer countries (Agarwal and Gaule (2020)) affirmative action policies or other non purely meritocratic selection rules might cause improvements in the allocation of talent.

The findings of our paper join those of many other scholars that have shown how intergenerational mobility is affected by early life conditions (e.g. see the review in Heckman and Mosso (2014)). Educational policy informed by careful analysis can address inequities at early stages with policies such as funding redistribution (Biasi (2021)) and provision of incentives to guarantee quality teaching in all public schools (Bobba et al. (2021)). Our findings show once more that unaddressed upstream disparities will lead to further inequities down the road, frustrating well-meaning policies, and highlights the role that meritocratic criteria in college admissions play in amplifying them.

References

Abdulkadiroğlu, A., Pathak, P. A., Schellenberg, J., and Walters, C. R. (2020). Do Parents Value School Effectiveness? *American Economic Review*, 110(5):1502–39.

Agarwal, R. and Gaule, P. (2020). Invisible Geniuses: Could the Knowledge Frontier Advance Faster? *American Economic Review: Insights*, 2(4):409–24.

Akbarpour, M., Dworczak, P., Duke Kominers, S., Alves, M., Ely, J., Jagadeesan, R., Kahn, A., Loertscher, S., Muir, E., Pavan, A., Stantcheva, S., and Strulovici, B. (2020). Redistributive Allocation Mechanisms *.

Alderman, H., Orazem, P. F., and Paterno, E. M. (2001). School quality, school cost, and the public/private school of choices of low-income households in Pakistan. *Journal of Human Resources*, 36(2):304–326.

Arcidiacono, P., Kinsler, J., and Ransom, T. (2021). Legacy and Athlete Preferences at Harvard. *https://doi.org/10.1086/713744*.

Arcidiacono, P. and Lovenheim, M. (2016). Affirmative action and the quality-fit trade-off.

Bailey, M. and Dynarski, S. (2011). Gains and Gaps: Changing Inequality in U.S. College Entry and Completion. *National Bureau of Economic Research*.

Baker, A., Larcker, D. F., and Wang, C. C. Y. (2021). How Much Should We Trust Staggered Difference-In-Differences Estimates? *SSRN Electronic Journal*.

Barro, R. J. (1991). Economic Growth in a Cross Section of Countries. *The Quarterly Journal of Economics*, 106(2):407–443.

Bautista, M. A., González, F., Martinez, L. R., Munoz, P., and Prem, M. (2020). Dictatorship, Higher Education and Social Mobility. *SSRN Electronic Journal*.

Bedard, K. (2001). Human capital versus signaling models: University access and high school dropouts. *Journal of Political Economy*, 109(4):749–775.

Belskaya, V., Peter, K. S., and Posso, C. M. (2020). Heterogeneity in the effect of college expansion policy on wages: Evidence from the Russian labor market. *Journal of Human Capital*, 14(1):84–121.

Beuermann, D., Jackson, C. K., Navarro-Sola, L., and Pardo, F. (2018). What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output. *National Bureau of Economic Research Working Paper Series*.

Bianchi, N. (2020). The Indirect Effects of Educational Expansions: Evidence from a Large Enrollment Increase in University Majors. *https://doi.org/10.1086/706050*, 38(3):767–804.

Biasi, B. (2021). School Finance Equalization Increases Intergenerational Mobility *.

Black, S., Denning, J., and Rothstein, J. (2020). Winners and Losers? The Effect of Gaining and Losing Access to Selective Colleges on Education and Labor Market Outcomes. Technical report, National Bureau of Economic Research, Cambridge, MA.

Bleemer, Z., Brady, H., Card, D., Cummins, J., Delong, B., Goodman, J., Kline, P., Moretti, E., Olney, M., Rothstein, J., Steier, Z., and Walters, C. (2020). Top Percent Policies and the Return to Postsecondary Selectivity *. *SSRN*, (December).

Bobba, M., Ederer, T., Leon-Ciliotta, G., Neilson, C., Nieddu, M., and Bobba Tim Ederer Gianmarco León-Ciliotta Christopher Neilson Marco Nieddu, M. A. (2021). "Teacher Compensation and Structural Inequality: Evidence from Centralized Teacher School Choice in Peru" Teacher Compensation and Structural Inequality: Evidence from Centralized Teacher School Choice in Perú.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting Event Study Designs: Robust and Efficienct Estimation. *Work in Progress*, pages 1–48.

Callaway, B. and Sant'Anna, P. H. (2020). Difference-in-Differences with multiple time periods. *Journal of Econometrics*.

Card, D. (1993). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. Technical report, National Bureau of Economic Research, Cambridge, MA.

Carrell, S. and Sacerdote, B. (2017). Why do college-going interventions work? *American Economic Journal: Applied Economics*, 9(3):124–151.

Chatterjee, C., Hanushek, E., and Mahendiran, S. (2020). Can Greater Access to Education Be Inequitable? New Evidence from India's Right to Education Act. Technical report, National Bureau of Economic Research, Cambridge, MA.

Chetty, R., Friedman, J. N., Saez, E., Turner, N., and Yagan, D. (2020). Income Segregation and Intergenerational Mobility across Colleges in the United States. *Quarterly Journal of Economics*, 135(3):1567–1633.

Cuenca, R. (2015). La educación universitaria en el Perú : democracia, expansión y desigualdades. *Instituto de Estudios Peruanos*, 4(1850):53–74.

Currie, J. and Moretti, E. (2003). Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings. *The Quarterly Journal of Economics*, 118(4):1495–1532.

Desa, U. N. and others (2016). Transforming our world: The 2030 agenda for sustainable development.

Dinerstein, M. and Smith, T. (2014). Quantifying the Supply Response of Private Schools to Public Policies.

Duflo, E., Dupas, P., and Kremer, M. (2015). Education, HIV, and Early Fertility: Experimental Evidence from Kenya. *American Economic Review*, 105(9):2757–97.

Eisenhauer, P., Heckman, J. J., and Mosso, S. (2015). Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments. *International Economic Review*, 56(2):331–357.

Ferreyra, M. M., Avitabile, C., Botero Álvarez, J., Haimovich Paz, F., and Urzúa, S. (2017). *At a Crossroads: Higher Education in Latin America and the Caribbean*. World Bank, Washington, DC.

Fuller, W. C., Manski, C. F., and Wise, D. A. (1982). New Evidence on the Economic Determinants of Postsecondary Schooling Choices. *The Journal of Human Resources*, 17(4):477.

Gelber, A. and Isen, A. (2013). Children's schooling and parents' behavior: Evidence from the Head Start Impact Study. *Journal of Public Economics*, 101(1):25–38.

Gennaioli, N., La Porta, R., Lopez-de Silanes, F., and Shleifer, A. (2013). Human Capital and Regional Development. *The Quarterly Journal of Economics*, 128(1):105–164.

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

Hanushek, E., Peterson, P., Talpey, L., and Woessmann, L. (2020). Long-run Trends in the U.S. SES-Achievement Gap. Technical report, National Bureau of Economic Research, Cambridge, MA.

Hastings, J., Hortaçsu, A., and Syverson, C. (2017). Sales Force and Competition in Financial Product Markets: The Case of Mexico's Social Security Privatization. *Econometrica*, 85(6):1723–1761.

Hastings, J. S., Neilson, C. A., Ramirez, A., and Zimmerman, S. D. (2016). (Un)informed college and major choice: Evidence from linked survey and administrative data. *Economics of Education Review*, 51:136–151.

Heckman, J. J. and Mosso, S. (2014). The Economics of Human Development and Social Mobility. *http://dx.doi.org/10.1146/annurev-economics-080213-040753*, 6:689–733.

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The Allocation of Talent and U.S. Economic Growth. *Econometrica*, 87(5):1439–1474.

Jagnani, M. and Khanna, G. (2020). The effects of elite public colleges on primary and secondary schooling markets in India. *Journal of Development Economics*, 146:102512.

Kapor, A., Neilson, C. A., Zimmerman, S. D., Clark, W., Davis-Googe, S., Ross-Lee, C., Pathak, P., Ferreyra, M. M., Agarwal, N., Budish, E., and Fu, C. (2017). Heterogeneous Beliefs and School Choice Mechanisms.

Kyui, N. (2016). Expansion of higher education, employment and wages: Evidence from the Russian Transition. *Labour Economics*, 39:68–87.

Mello, U. (2019). Centralized Admissions, Affirmative Action and Access of Low-income Students to Higher Education. *American Economic Journal: Economic Policy*.

Ministerio de Educación (2020). Política Nacional de Educación Superior y Técnico-Productiva.

Mokyr, J. (2005). Long-Term Economic Growth and the History of Technology. *Handbook of Economic Growth*, 1(SUPPL. PART B):1113–1180.

Neilson, C. (2013). Targeted Vouchers, Competition Among Schools, and the Academic Achievement of Poor Students. *Job Market Paper*.

OECD (2016). *Avanzando hacia una mejor educacion para Perú*, volume 3.

Oppedisano, V. (2011). The (adverse) effects of expanding higher education: Evidence from Italy. *Economics of Education Review*, 30(5):997–1008.

Osili, U. O. and Long, B. T. (2008). Does female schooling reduce fertility? Evidence from Nigeria. *Journal of Development Economics*, 87(1):57–75.

Pop-Eleches, C. and Urquiola, M. (2013). Going to a Better School: Effects and Behavioral Responses. *American Economic Review*, 103(4):1289–1324.

Russell, L., Yu, L., and Andrews, M. J. (2021). Historical Happenstance and Local Educational Attainment: Evidence from the Establishment of U.S. Colleges. Technical report.

Sánchez, A., Favara, M., and Porter, C. (2021). Stratification of returns to higher education in Peru: the role of education quality and major choices Stratification of returns to higher education in Peru: the role of education quality and major choices 1. Technical report.

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

Squicciarini, M. P. and Voigtländer, N. (2015). Human Capital and Industrialization: Evidence from the Age of Enlightenment. *The Quarterly Journal of Economics*, 130(4):1825–1883.

Stadtman, V. A. (1970). *The University of California, 1868–1968*. McGraw-Hill, New York.

Sun, L. and Abraham, S. (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

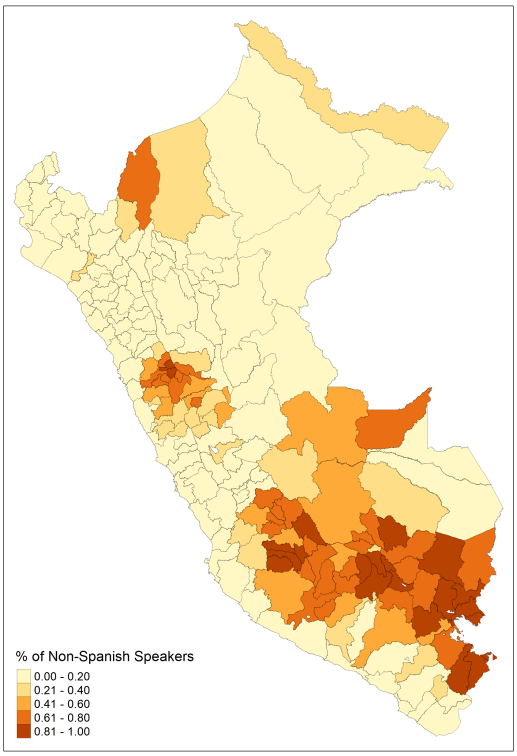
SUNEDU (2020). II Informe Bienal sobre la Realidad Universitaria en el Perú | Gobierno del Perú. Technical report.

UNESCO (2020). Global education monitoring report, 2020, Latin America and the Caribbean: inclusion and education: all means all.

A Additional figures

Figure 10: Dimensions and Ethnic Dispersion of Peru

Dimensions of Peru and California



Native Language Speakers (%)

Figure 11: Distribution of Treatment

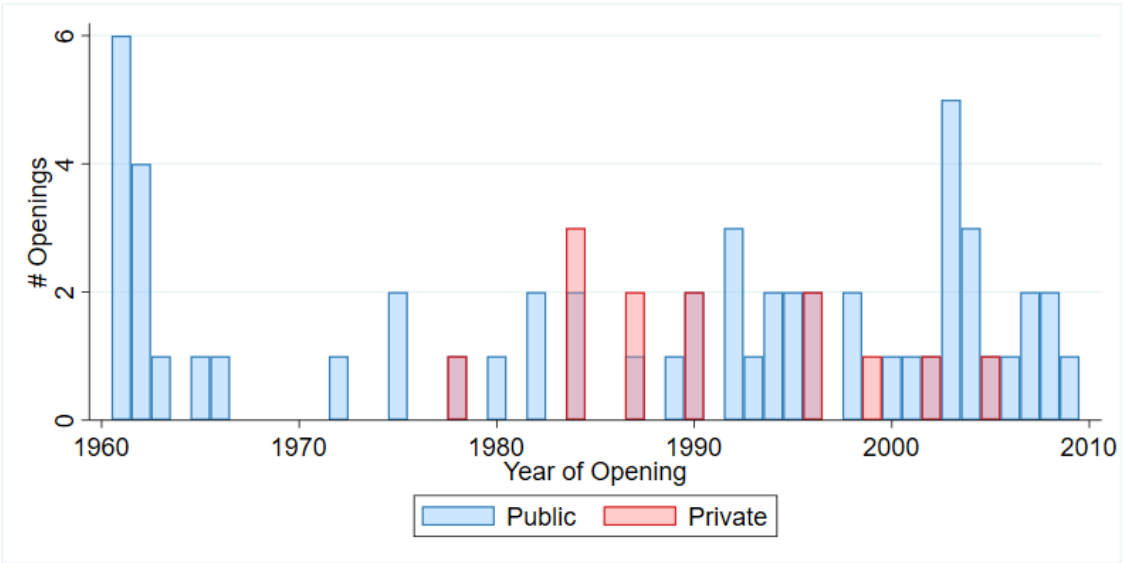


Figure 12: Time Between High School Graduation and College Enrollment

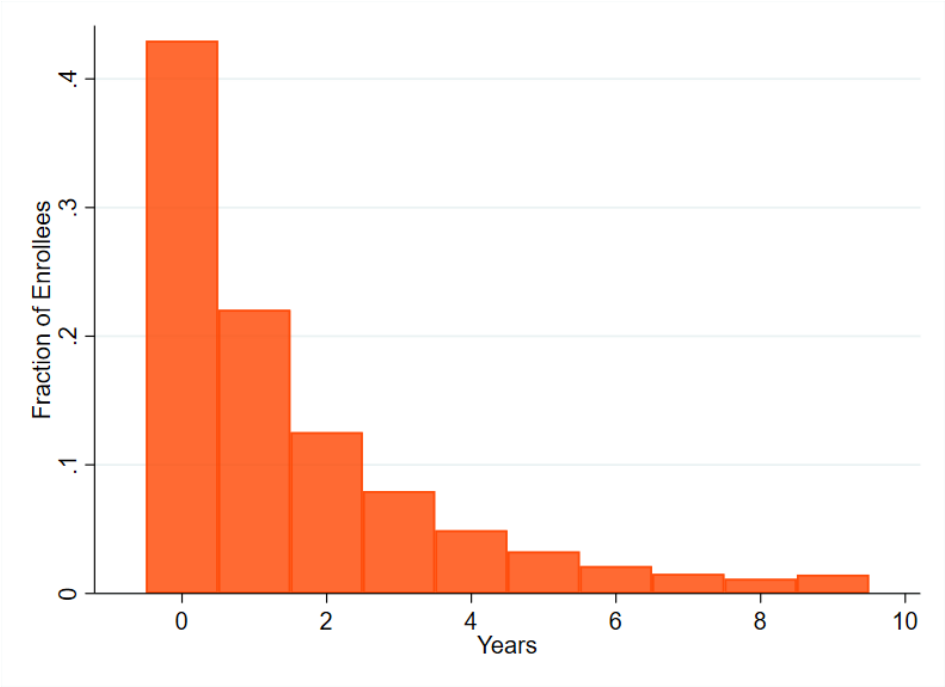
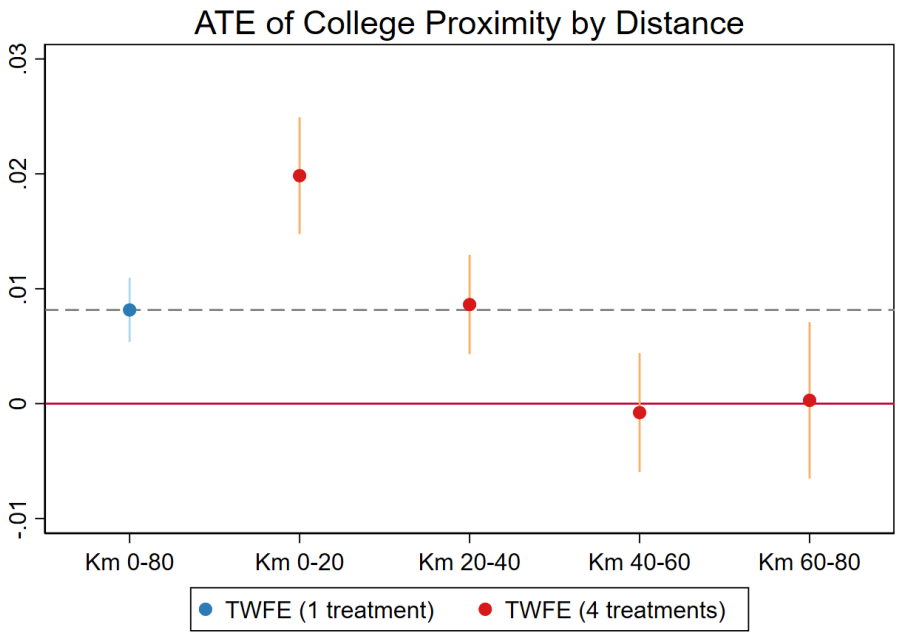


Figure 13: Effects by distance bins



[OTHER FIGURES AVAILABLE UPON REQUEST]

B Additional Tables

Table 7: Replicating Table 2 – Distance as Treatment

	College Enrollment	College Completion
TWFE		
	0.0256***	0.0212***
	(0.00550)	(0.00462)
All Cohorts		
	.0121**	.0093*
	(.0061)	(.0053)
School Age		
	.0101***	.0083***
	(.0027)	(.0025)

Table 8: Replicating Table 3 – Distance as Treatment

	College Enrollment	College Completion
By Ethnicity		
Majority	.0141*	.0102*
	(.0076)	(.0061)
Minority	-.0008	.0017
	(.013)	(.013)
By Gender		
Female	.0152**	.0108*
	(.0068)	(.0062)
Male	.005	.0048
	(.0067)	(.0061)

[OTHER TABLES AVAILABLE UPON REQUEST]

C Students' Mobility

[AVAILABLE UPON REQUEST]

D Building dataset for college openings

Data on the creation of new universities is created by starting from an excel file provided by the Ministry of Education. From this file we assume that if a university offers a major in a certain province/district, then it has a campus there. With this list of campuses associated to locations, we proceed to find the date of opening of such university. We refer to several sources, in an attempt to find the most credible year in which such university admitted students:

- The first source of information is the date when the first major was registered in a given province. This is most likely giving us an upper bound on the real date of opening of the university/campus, as we observe that in most cases for which we find reliable information contradicting the date obtained through this method the most credible estimate is *pre-dating* major registration.

- We use information obtained from universities' websites. This information is generally displayed in the "History" and "Transparency" [65] sections of websites. When information on the year that the first cohort was admitted is available, this is considered the most credible source of information. Unfortunately, in many cases only information on the year of creation (which can differ substantially from the year of first admission) is included. In some other cases, information is available for the first campus of the university, but not for more recent ones.

- In 2010 a national census of university students was performed: this includes information about where students are enrolled and the year they first enrolled. We use this information to validate other sources, checking if students enrolled at a given campus earlier than we thought it opened its doors. A sharp increase (starting from 0) in enrollees in one specific year can also signal that the university started admitting students in that year.

- When the former sources of information were not enough, we used information published by other government entities. These include announcements issued by SUNEDU

[65] A section that holds the university administrative documents and the statistics of admitted, enrolled and graduated students

denying or granting institutional licensing, information of public investment projects released by the Ministry of Economy and Finance, and statistics about college enrollment published by the National Institute of Statistics and Information (INEI).

- News from local newspapers is also used for the cases where no other information is available to infer the date of first admission. This includes reports of the university holding admission exams or a journalistic reconstruction of a university's history.

Through this process, we obtain the year that the university was actively teaching a cohort of students. To account for the fact that admissions start earlier, influencing behavior of students in the previous year, we will consider the year before for our shock. As an example, if college A had its first cohort enrolled in 2000 we will consider the province where it opened as shock starting in 1999.[66]

We obtain 70 provinces that had a university open in the period 1960 to 2010. These provinces represent our sample of interest.

Figure 1 shows the geographical dispersion of the provinces in our sample.

[66]Notice that Peru being in the Southern Hemisphere, its academic year is aligned with the solar year with the first academic semester in the first 7 months of the year. Classes typically begin in March and finish in December.

E Building tuition data

We build the tuition data separately for public universities and for private ones.

E.1 Tuition data for public universities

The unique regular tariff that public universities charge per semester on students are the enrollment fees. There are two different types of enrollment fees:

- For incoming students
- For regular students: For those who are in their second semester or higher

These amounts can be found in the Unique Text of Administrative Procedures (or TUPA by the Spanish acronym) of each university. By the *Ley Universitaria*, universities are required to upload this document to their websites, under the section "University Transparency".

E.2 Tuition data for private universities

The regular tariffs that private universities charge per semester on students are the enrollment fees and tuition.

- Enrollment fees can be found in the Unique Text of Administrative Procedures of each university.
- Gathering the information of tuition required more steps. Most of the time, the section of "University transparency" contains it under the subsections "Unique Text of Administrative Procedure" or "List of payments required". In other cases, we had to resort to informative brochures available through the admission page of universities.
- The number of tuition payments per semester can be found along with the tuition or in the payment schedule of the university.

F TWFE specification and negative weighting

[AVAILABLE UPON REQUEST]

G Probability of Admission (γ) and Beliefs ($\tilde{\gamma}$)

Table 9: Predicting Probability of Admission, by Demographic Group

	(1)	(2)	(3)	(4)
High GPA	0.173***	0.173***	0.173***	0.173***
	(0.00867)	(0.00864)	(0.00853)	(0.00849)
Low GPA	-0.0583***	-0.0584***	-0.0596***	-0.0597***
	(0.00777)	(0.00772)	(0.00783)	(0.00778)
High Qual. HS	0.0523***	0.0669***	0.0490***	0.0627***
	(0.00466)	(0.00787)	(0.00494)	(0.00756)
Female	-0.0623***	-0.0650***	-0.0624***	-0.0651***
	(0.00491)	(0.00834)	(0.00490)	(0.00836)
Ethnic Min.	-0.0278***	-0.0188	-0.0102	0.0143
	(0.00749)	(0.0187)	(0.00967)	(0.0217)
High SES	0.0273***	0.0364***	0.0283***	0.0369***
	(0.00532)	(0.00909)	(0.00527)	(0.0114)
HS*Female		-0.00728		-0.00435
		(0.00739)		(0.00757)
HS*Minority		-0.0187		-0.0333**
		(0.0133)		(0.0140)
HS*SES		-0.0183**		-0.0179**
		(0.00819)		(0.00820)
Close*Female		0.00863		0.00655
		(0.00939)		(0.00944)
Close*Minority		-0.00229		-0.0230
		(0.0212)		(0.0230)
Close*SES		0.00283		0.00244
		(0.0102)		(0.0119)
College Close			0.0446***	0.0452***
			(0.0111)	(0.0146)
Province FE	Yes	Yes	No	No
Region FE	No	No	Yes	Yes
Lima FE	No	No	Yes	Yes
Observations	117233	117233	117234	117234
R^2	0.805	0.806	0.731	0.732

Table 10: Linear Probability Model for Probability of Admission

	(1) Admission Probability
High GPA	0.173***
	(0.00849)
Low GPA	-0.0597***
	(0.00778)
High Qual. HS	0.0627***
	(0.00756)
High SES	0.0369***
	(0.0114)
Ethnic Min.	0.0143
	(0.0217)
Female	-0.0651***
	(0.00836)
High Qual. HS × High SES	-0.0179**
	(0.00820)
High Qual. HS × Ethnic Min.	-0.0333**
	(0.0140)
High Qual. HS × Female	-0.00435
	(0.00757)
College Close	0.0452***
	(0.0146)
College Close × High SES	0.00244
	(0.0119)
College Close × Ethnic Min.	-0.0230
	(0.0230)
College Close × Female	0.00655
	(0.00944)
Observations	117234

Table 11: Beliefs and Observed Probabilities of Admission

	(1) Beliefs	(2) Probability
High GPA	0.0944***	0.173***
	(0.0214)	(0.00853)
Low GPA	-0.211***	-0.0596***
	(0.0449)	(0.00783)
High Qual. HS	0.149***	0.0490***
	(0.0114)	(0.00494)
High SES	0.0983***	0.0283***
	(0.0241)	(0.00527)
Ethnic Min.	-0.0446**	-0.0102
	(0.0217)	(0.00967)
Female	0.0434***	-0.0624***
	(0.0131)	(0.00490)
College Close	0.00792	0.0446***
	(0.0184)	(0.0111)
Observations	459286	117234

H Likelihood Function

Define pairs $\omega \equiv (j, a)$ where $j \in J = \{pub, priv, hs, do\}$ is the realized outcome and $a \in A = \{0, 1\}$ representing whether the student applied to a public university or not.[67] The likelihood function will also depend on admission probability γ_i.[68]

The probability of observing the pair ω for individual i in cell $m(i)$ is:

$$P_i = Pr(\omega = (j_i, a_i)|m(i)) = \begin{cases} (1-\Gamma) & if \quad \omega = (do, .) \\ \Gamma * \gamma * Pr(u_{pub} > u_{priv} \quad \& \quad u_{pub} > u_{hs}) & if \quad \omega = (pub, 1) \\ \Gamma * (1-\gamma) * Pr(u_{pub} > u_{priv} > u_{work}) & if \quad \omega = (priv, 1) \\ \Gamma * (1-\gamma) * Pr(u_{pub} > u_{work} > u_{priv}) & if \quad \omega = (hs, 1) \\ \Gamma * Pr(u_{priv} > u_{pub} \quad \& \quad u_{priv} > u_{hs}) & if \quad \omega = (priv, 0) \\ \Gamma * Pr(u_{hs} > u_{pub} \quad \& \quad u_{hs} > u_{priv}) & if \quad \omega = (hs, 0) \end{cases}$$

where $(1-\Gamma) \equiv Pr(E[max\{u_{hs}, u_{priv}, \tilde{\gamma} u_{pub} + (1-\tilde{\gamma}) max\{u_{hs}, u_{priv}\}\}] < 0)$ and we omit the index i on the right-hand side to reduce clutter.

The likelihood function will then be equal to $\mathcal{L}(X, \theta) = \prod_i P_i$. In the first step, we estimate mean utilities for each option for each group $m(i)$: this is equivalent to estimating the model separately for each group with an intercept for each option (except for the normalization of u_{do}). During this step, beliefs about the probability of admission, $\tilde{\gamma}_{m(i)}$ are also backed out. We provide results using beliefs backed out within the MLE procedure and set equal to the predictions from the LPM described above, $\hat{\gamma}_i$. $\tilde{\gamma}_{m(i)}$ is identified by the parametric assumptions on the expected utility from graduation and by trying to rationalize dropout behavior.[69] Table 11 shows how the recovered beliefs relate with demographic characteristics, comparing them to the observed probabilities of admission. The results show that beliefs relate similarly to demographics as observed probability, except for gender: women appear to be overconfident regarding their chances of admission.

[67] Because students can't be admitted to public university if they don't apply, we can omit the indicator for the pair $(pub, 0)$. Similarly, students that drop out of high school cannot apply at any public college, so $\Pr(\omega = (do, 0)|m(i)) = Pr(\omega = (do, 1)|m(i))$.

[68] In the estimation, we will replace γ_i with $\hat{\gamma}_i$ estimated using the model described above.

[69] Intuitively, there are two decisions being made by everyone. The second one is based on the utility from each option. Taking these utilities as fixed, the first choice depends on the beliefs about the probability of admission.